SYNONYMOUS CODONS AFFECT POLYSOME SPACING, PROTEIN PRODUCTION AND PROTEIN FOLDING STRESS: STUDIES OF BACTERIAL TRANSLATION USING RIBOSOME PROFILING

by

Andrew Ted Martens

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

August, 2017

Abstract

The acquisition of protein secondary and tertiary structure depends on the primary sequence of amino acids. However, predicting a protein's folded structure is difficult even with the knowledge of its sequence. It has been suggested that, in addition to encoding the amino acid sequence, genes also encode kinetic information which regulates the ribosome's translation rate. This information might guide nascent protein folding during translation.

With the advent of ribosome profiling, a high-throughput sequencing technique which quantifies ribosome density on mRNA, it is now possible to investigate this hypothesis in greater detail. Here, a new way to analyze ribosome profiling data is presented, confirming that ribosome profiling detects ribosome pauses at slow codons (chapter 2). This method is able to precisely determine the locations of the ribosome aminoacyl and peptidyl transfer sites within the ribosome footprint. Next, a simulation tool which models the progression of ribo-

ABSTRACT

somes along an mRNA is used to explore the effects of translation initiation and elongation rates on protein expression (chapter 3). This tool can be used to generate testable predictions for how changing the translation rate should affect various experimental observables, including ribosome density. New experimental data, collected from the bacterium *Escherichia coli*, demonstrate that the sequence of the Firefly (*Photinus pyralis*) Luciferase mRNA affects its ribosome occupancy (chapter 4). Importantly, ribosome occupancy is differentially influenced by synonymous codons.

These data also show that Luc expression is controlled by the 15 codons immediately downstream of the start codon and that greater Luciferase expression levels progressively activate the "heat shock" response (chapter 5). However, this response appears to saturate, suggesting that the overexpression of foreign proteins in *E. coli* readily overwhelms the endogenous chaperone system. This result demonstrates that expression level, rather than translation kinetics, determines the yield of folded Luciferase protein in *E. coli*.

Primary Reader: Dr. Vincent J. Hilser

Secondary Reader: Dr. Rachel Green

Acknowledgments

I have received much help and support during my graduate studies. My mentor, Vince, has given me the freedom to explore my own ideas. He has also instilled in me the importance of carrying out science driven by quantitative theory and experimentation. Likewise, Bob Schleif and Elijah Roberts have provided me with excellent guidance while giving much needed advice and feedback. Rachel Green's expertise and input at the later stages has been extremely helpful, and the quality of this dissertation is much improved as a result.

People outside of my committee also provided assistance. Nick Ingolia helped convince me to do ribosome profiling, despite how daunting it seemed. In retrospect, using this technique has been an excellent learning experience. James Taylor's input on my first paper, and Christian Kaiser's feedback over the years, have been invaluable. I also thank my undergraduate advisor, Malcolm Campbell, for teaching me many years ago that a student's job is to learn. This advice has helped me keep any

setbacks in perspective.

The biology and biophysics community has been very generous in sharing time, ideas and equipment. Many of the experiments presented here would not have been possible without this help. In particular, I thank Bob Horner and the Beemon, Chen, Kaiser, Schleif, Schroer, Van Doren, Wendland, Woodson and Zappulla labs.

The other members of the Hilser lab, past and present, have been fun to be around, but also helpful and intellectually engaging. Thank you for making work a pleasant experience for the last 6 years. Outside of work my many friends, both in Baltimore and beyond, have been great fun, and spending time with them has taken the edge off of the graduate school grind.

Thanks, Pops, for telling me about the importance of persistence. Thanks mom for proofreading drafts of this very long document and for your encouragement. Thanks Cynthia and Sasha for being my lifelong role models. Finally, and most importantly, I thank my family for being there for me.

Dedication

This dissertation is dedicated in memory of Ted F. Martens and Michael F. Martens.

Contents

Tables

Figures

Chapter 1

Introduction

All proteins are composed of the same amino acid residue types, differing in number and order, yet there are at least 1,393 stable folds [Berman et al., 2000] (Fig. 1.1, p. 2). A major unresolved question in protein biophysics is how these amino acid sequences give rise to biologically functional, folded proteins. Perturbations to protein folding, either through mutation or from environmental insults, can cause cellular stress and are implicated in disease. Yet, it remains extremely difficult to predict a protein fold from its primary sequence [Dill et al., 2008]. Greater knowledge of the protein folding process will inform both fundamental biophysical principles as well as medical approaches to such diseases, and the discovery of new biological mechanisms which can promote or inhibit these processes is of great interest.

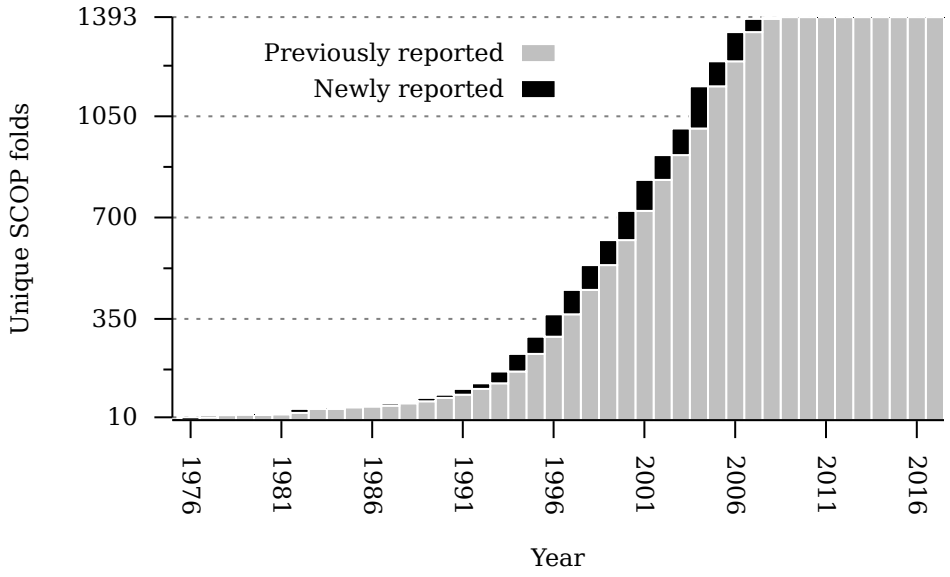


Figure 1.1: Proteins can adopt many kinds of folds. The Structural Classification of Proteins database reports 1,393 known unique folds. This diversity of folds has emerged from many amino acid sequences, predominantly composed of the same twenty amino acids, yet predicting a fold from such a sequence remains a challenge. Data are from the Protein Data Bank.

Classic experiments demonstrated how a protein's biological "native" state is typically the lowest energy state [Haber and Anfinsen, 1962]. However, proteins cannot sample all possible conformations [Levinthal, 1969]. It has been suggested that proteins have folding pathways which constrain the sampling of conformational space, either during or after synthesis on the ribosome [Levinthal, 1968]. Since a nascent polypeptide emerges from the ribosome at the same rate as the ribosome translocates the mRNA, slower translation could provide more time for a newly-emerged domain to fold in isolation, before other domains emerge. Conversely, faster translation could prevent misfolding by rapidly providing a nascent domain with new amino acids and steering the protein towards the native fold. Such kinetic information could help guide protein folding predictions [Ellis et al., 2010, Trovato and O'Brien, 2016].

It has been suggested that life has evolved ways to synchronize the emergence of a nascent protein with its co-translational folding. In particular, this information could be genetically encoded by specific sequences of synonymous codons which have different translation elongation rates. Protein secondary structural elements, or boundaries between elements, could occur in conjunction with translation pause sites [Purvis et al., 1987, Krasheninnikov et al., 1989]. This mechanism re-

quires that proteins fold on the ribosome (Ch. 1.7, p. 34) and that the translation rate is genetically flexible (Ch. 1.6, p. 28). Understanding this relationship therefore depends on accurately detecting changes in translation rates (Ch. 1.5, p. 24) and corresponding changes in co-translational protein folding (Ch. 1.7).

Currently both aspects of this hypothetical mechanism, a regulated translation rate synchronized with protein folding on the ribosome, have been difficult to generalize. In particular, precise and accurate position-specific measurement of the *in vivo* translation rate has not yet been achieved. However, a more recent experimental method which could address this issue has emerged. The high-throughput sequencing of ribosome-bound mRNA fragments, known as ribosome profiling [Ingolia et al., 2009], measures the relative ribosome occupancy of mRNA. If the occupancy of a particular codon is inversely proportional to the ribosome's transit time, then this method could provide the necessary information relating the mRNA sequence with its translation rate (Ch. 1.8, p. 38). This dissertation explores the use of ribosome profiling to quantify relative translation rates and uses a model protein system, Firefly Luciferase (Luc), to investigate the relationship between the translation rate and co-translational protein folding in the bacterium *Escherichia coli*.

1.1 Synonymous codon frequencies in genes and genomes are not random

As the genetic code was deciphered it was found to be degenerate, containing synonymous codons which encode the same amino acid [Khorana, 1968]. Although these codons are nominally equivalent, regulatory codons with alternative translation rates were proposed as a mechanism for gene regulation [Itano, 1957, Ames and Hartman, 1963, Stent, 1964, Itano, 1965, Itano, 1968]. Yet neither the biological significance of codon degeneracy nor its extent were clear until genes and genomes were sequenced. Here is a brief history of key developments in this field.

In 1969 the first mRNA sequence, from bacteriophage MS2, was chemically determined [Adams et al., 1969]. The sequence was 57 nucleotides long and contained synonymous codons. A larger RNA fragment, encoding the MS2 coat protein, was sequenced and found to encode a hairpin [Jou et al., 1971, Hasegawa et al., 1979]. It was noted that multiple RNA secondary structures were compatible with the same amino acid sequence due to the degeneracy of the genetic code. When

the complete sequence of the MS2 A-protein gene was determined (392 codons), it was found that some synonymous codons were significantly more frequent than others [Fiers et al., 1975]. This result was replicated with gene F of bacteriophage φX174 [Air et al., 1976] and rabbit beta [Efstradiatis et al., 1977] and alpha [Heindell et al., 1978] globin genes. These experiments demonstrated that synonymous codons naturally occur within the same gene.

It was subsequently shown that groups of genes within a genome can have distinctive codon patterns. Sequences of highly-expressed *E. coli* genes were shown to have skewed frequency statistics [Post et al., 1979], as were several yeast genes [Bennetzen and Hall, 1982]. These statistics differed from lowly-expressed genes [Farabaugh, 1978]. Studies in yeast then showed how to predict gene expression levels from their codon composition [Sharp et al., 1986]. Similar results have been found in *E. coli* and in *Bacillus subtilis* [Moszer et al., 1999]. Besides highly and lowly expressed genes, a third category of genes related to outer membrane and lipopolysaccharide (LPS) genes was also discovered [Médigue et al., 1991, Guerdoux-Jamet et al., 1997].

Curiously, it is often the case that codon frequencies within an organism, between organisms, or even within a single mRNA are not uniform. For example, the MS2 coat protein mRNA resembles typical *E. coli*

mRNAs, but the MS2 A protein mRNA does not [Elton et al., 1976]. This fact contradicts the observation that the mRNAs from a given organism cluster together because they have similar synonymous codon frequencies, but do not cluster with mRNAs from other organisms [Grantham et al., 1980]. Intriguingly, the codon frequencies within the same immunoglobin, γB-crystallin or silk fibroin mRNAs are unevenly distributed [Miyata et al., 1979, Komar and Jaenicke, 1995, Chavancy et al., 1979, Mita et al., 1988]. Constraints on the ordering, or context, of codons have also been demonstrated [Lipman and Wilbur, 1983]. These observations suggest the possibility that synonymous codons are not necessarily functionally identical.

1.2 Causes and effects of codon identity

A compelling explanation for biased codon frequencies was proposed soon after their discovery. Quantification of the genomic tRNA gene copy numbers and cellular tRNA abundances led to the hypothesis that the codon frequencies within a genome have co-evolved with the available tRNAs [Ikemura, 1981b, Ikemura, 1981a, Ikemura, 1982, Ikemura

and Ozeki, 1983, Ikemura, 1985]. This effect may be the result of selection against wobble recognition [Fitch, 1976], or in favor of moderate interaction energies [Grosjean and Fiers, 1982]. This idea was extended to adaptation within specialized tissues, such as the silk worm (*Bombyx mori*) silk gland [Chavancy et al., 1975, Garel, 1976, Chavancy and Fournier, 1979, Chevallier and Garel, 1979, Chavancy et al., 1981, Garel et al., 1981] and in reticulocytes [Hatfield et al., 1982].

To compare genes in terms of their codon repertoires, normalized scales have been developed to rank the codons. Such rankings include codon frequency [Ikemura, 1981a, Bennetzen and Hall, 1982]; stochastic searches [Varenne et al., 1984, Zhang et al., 2009]; codon adaptation (**c**odon **A**daptation **I**ndex, [Sharp and Li, 1987]); tRNA adaptation (**t**RNA **A**daptation **I**ndex; [dos Reis et al., 2004] and a similar index [Spencer et al., 2012], Table 4.1, p. 145), relative frequencies (%MinMax, [Clarke and Clark, 2008]), and "supply and demand" models of tRNA and codons [Pechmann and Frydman, 2013]. These types of rankings have proven popular for generating an expectation for both a gene's overall codon composition and for local fluctuations within.

The correlations between codon frequencies, tRNA gene copy number and tRNA concentration are substantial. However, these relationships are not definitive. For instance, tRNA gene expression in silk

glands is regulated [Chevallier and Garel, 1982], with some tRNA genes transcribed at higher rates than others. In addition, the tRNA content within *E. coli* cells adapts to the growth rate [Dong et al., 1996], yet cells are unable to adapt to the over-expression of a gene with uncommon codons [Dong et al., 1995]. These results suggest that cells might have evolved mechanisms which can match tRNA content to specific growth conditions, but do not necessarily respond *per se* to the mRNA burden. tRNA gene number and genomic codon frequencies are often, but not always, related.

Sometimes, special codons sense nutrient conditions. For example, the tryptophan operon in *E. coli* is regulated in response to $tRNA^{Trp}$ availability. The ribosome slows down while translating a pair of tryptophan codons, but only if $tRNA^{Trp}$ is limiting [Yanofsky, 1981]. It was also shown that 70% of $tRNA^{Leu}$ in *E. coli* cell is aminoacylated, but that this value drops below 20% in the absence of leucine [Yegian et al., 1966]. Thus even if a tRNA is available, it might not be charged with an amino acid. More recently, high-throughput methods using microarrays have confirmed the general effects of amino acid deprivation on tRNA aminoacylation [Dittmar et al., 2005]. Importantly, models of tRNA aminoacylation states have suggested that certain codons are positioned within mRNAs to optimally respond to amino acid starvation [Elf et al., 2003].

These observations indicate that some codons can modulate the translation rate in response to environmental signals.

In fact, there are other potential causes and effects of non-random codon frequencies (Fig. 1.2, p. 11). Many of these are biologically important but not necessarily related to translation [Plotkin and Kudla, 2011, Pop et al., 2014]. For example, it has been suggested that tRNA and mRNA have co-evolved [Bulmer, 1987, Bulmer, 1991], that codons encode binding sites on DNA [Thurman et al., 2012] or RNA [Stergachis et al., 2013], that codons differentially regulate during the cell cycle [Frenkel-Morgenstern et al., 2012, Zhou et al., 2013], and that selection upon secondary structure [Mita et al., 1988] or GC content [Andersson et al., 1984, Palidwor et al., 2010, Babbitt et al., 2014, Newman et al., 2016] drives codon composition. The sheer number of possible effects of biased codon frequencies has made it difficult to unambiguously address causality. Nevertheless, it has now been clearly demonstrated, both theoretically and experimentally, that the translation elongation rate depends on the codon.

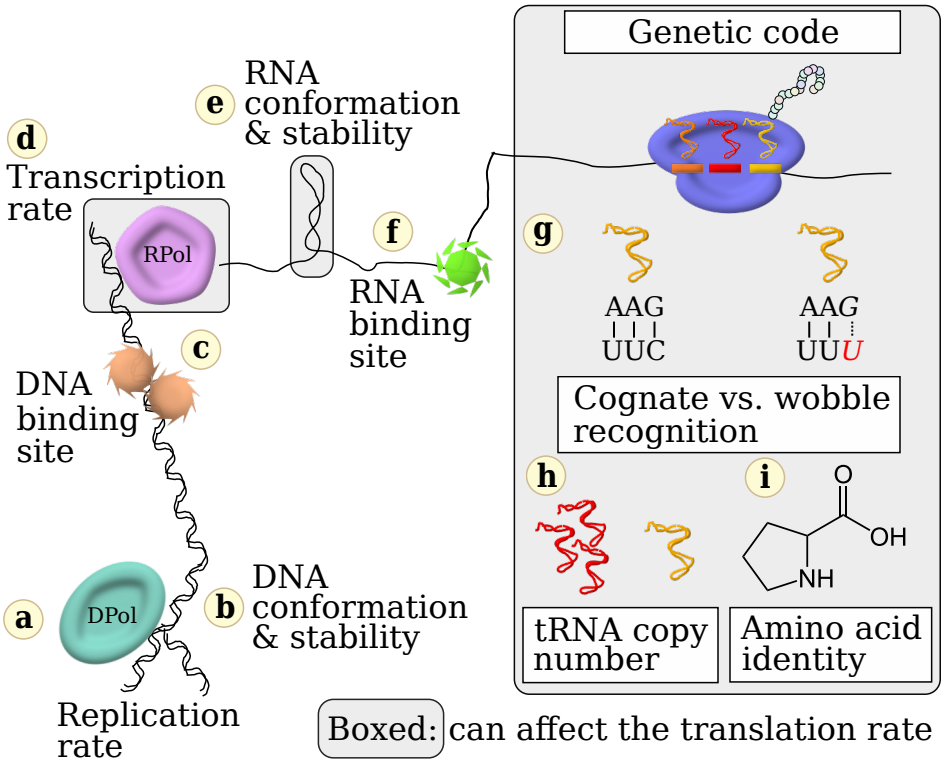


Figure 1.2: Synonymous changes can have many downstream effects. Replacing one codon with another in the mRNA requires making the equivalent change in the corresponding DNA. Altering a DNA sequence can have many consequences, only some of which have been shown to affect translation rates. a. the synthesis of a new DNA strand b. DNA conformation and stability c. binding sites on the DNA d. transcription rate e. RNA conformation and stability f. binding sites on the RNA g. Watson-Crick or wobble recognition h. copy number of the tRNAs capable of de-coding a given codon i. the incorporation of a new amino acid into an elongating polypeptide chain.

1.3 Mechanisms of codon-dependent elongation rates

No process can be perfectly accurate [Galas et al., 1986]. Consequently, non-equilibrium mechanisms have evolved which enhance accuracy at the expense of energy (such as ATP or GTP hydrolysis). Translation proceeds as a series of recognition events whose overall accuracy depends on the faithful aminoacylation of tRNAs by synthetases and the discrimination amongst tRNAs by a ribosome programmed with an mRNA. Kinetic models of proofreading during translation invoke schemes which directly integrate the accuracy of amino acid incorporation into a nascent polypeptide with the translation elongation rate (reviewed by [Zaher and Green, 2009]). Therefore, these accuracy mechanisms can explain how codons determine translation elongation rates. The following brief overview of ribosomal accuracy elaborates on this relationship.

Kinetic proofreading schemes in translation modulate the time required to process a codon. Two similar models explain how this duration can be lengthened. In one case, codon:anti-codon recognition is repeatedly evaluated, and mismatching tRNA are preferentially discarded [Hopfield, 1974]. Or, the “sticking time” of the anti-codon is

assessed, and any tRNA which fails to incorporate rapidly is rejected [Ninio, 1975]. A "double-trigger" mechanism which brings together both forms of proofreading has even been proposed [Ninio, 2006]. Crucially, these schemes couple proofreading with GTP hydrolysis, keeping the system far from equilibrium [Kurland, 1978]. Ribosome accuracy therefore depends on how the ribosome modulates the duration of the steps of an elongation cycle in response to inputs, such as the tRNA and mRNA, under non-equilibrium conditions (reviewed by [Rodnina and Wintermeyer, 2001]).

Translation elongation is the phase during which a ribosome repeatedly cycles through a stereotyped process which begins once a 70S ribosome forms. When the 30S (small) and 50S (large) ribosome subunits associate they establish three spaces within a cavity: the aminoacyl (A), peptidyl-transfer (P) and Exit (E) sites. An additional site (T), where EF-Tu (**E**longation **F**actor-**T**hermally **u**nstable) associates, remains outside the cavity. tRNA molecules are oriented within the cavity such that the anti-codon interfaces with the codon in the small subunit, while the aminoacyl 3′ end sits near the polypeptide chain exit tunnel in the large subunit. During translation the tRNAs transit through these three sites in a step-wise progression, A→P→E. Although the ribosome progresses one codon at a time, tRNAs transiently occupy hybrid states in which

one part has advanced relative to the small subunit. These states are denoted A/T, A/P, and P/E, indicating the positioning of the anti-codon and CCA arms, respectively [Moazed and Noller, 1989]. The translation elongation rate, also referred to as the ribosomal step time or transit time, is thus the time required for completing this cycle once, adding a single amino acid to the nascent polypeptide chain and moving the mRNA through the ribosome by one codon.

The genetic code is degenerate, therefore synonymous codons might have different translation rates. For example, two synonymous codons might encode the same amino acid but be recognized by a different tRNA. Such codons could be processed at different rates if the cognate tRNAs have different modifications, aminoacylation levels or concentrations [Ames and Hartman, 1963, Stent, 1964, Itano, 1965]. However, research into the determinants of ribosomal accuracy has shown how other steps in translation also have sequence-dependent rates. The relationship between accuracy and kinetics is not straightforward and depends on the details of codon:anti-codon recognition.

The wobble model of codon:anti-codon recognition states that the first two positions in codons must obey Watson-Crick base-pairing rules, but that alternative (wobble) pairings are allowed at the third position [Crick, 1966]. This model explains how a ribosome programmed with

an mRNA can discriminate between competing tRNAs through purely structural means — codon:anti-codon pairs which do not sterically match are always rejected. Yet this model also allows structural flexibility, accounting for how some synonymous codons are decoded by the same tRNAs. As a result, these rules still serve as good first approximations in understanding codon:anti-codon recognition [Ninio, 1973].

Yet, many anti-codons which recognize different codons are still similar, and mis-recognition events do happen [Kramer and Farabaugh, 2007, Kramer et al., 2010]. Remarkably, mutations outside the anti-codon loop can also affect tRNA reading. This observation led to the development of the waggle hypothesis, which states that strain imposed upon the tRNA provides additional discrimination during translation [Yarus and Smith, 1995]. Consequently, the ability to accurately discriminate all 64 codons imposes stringent constraints on tRNA sequences, and no organism possesses a full complement of tRNA anti-codons, an idea dubbed the "missing triplet" hypothesis [Ninio, 1971]. The evolution of kinetic proofreading mechanisms has enhanced the fidelity of translation, overcoming these limits of structural recognition, at the expense of kinetic efficiency.

Translation elongation can be summarized as follows. Ternary complex (EF-Tu, GTP and an aa-tRNA) repeatedly diffuses to and from the

ribosome. This step is rapid, and tRNA transiently occupies the A/T hybrid state [Rodnina et al., 1996, Pape et al., 1998]. If the codon and anti-codon have poor complementarity, then association is disfavored, dissociation dominates and the tRNA is quickly rejected with high probability. However, an eventual recognition event with a favorable interaction (cognate or near-cognate) promotes hydrolysis of GTP, release of EF-Tu·GDP and the transition of the aa-tRNA from the A/T state to the A/A state. While the aa-tRNA sits in the A site, and before EF-Tu·GDP has been released, a kinetic proofreading step takes place. Cognate tRNA are rapidly accepted, whereas other tRNAs are delayed. Longer delays increase the chance that the tRNA is rejected, and so this step preferentially discards non-cognate tRNA and thereby enhances the specificity of translation.

This model of translation elongation suggests two sources of codon-specific translation rates. The first is the time spent "searching" for a cognate tRNA out of the cellular pool of competing tRNA species. Estimates have suggested that this step is highly influential [Gouy and Grantham, 1980, Chavancy and Garel, 1981, Curran and Yarus, 1989, Andersson and Kurland, 1990]) and can be altered by tRNA concentration changes [Anderson, 1969]. In particular, this effect is amplified following sequestration of tRNA within the ribosome [Varenne et al., 1984,

Varenne and Lazdunski, 1986, Bonekamp and Jensen, 1988, Varenne et al., 1989, Menez et al., 2000] (Ch. 1.4). The second source of variability comes from the time spent proofreading. It has been inferred from experiments, using hyper-accurate and inaccurate mutant ribosomes, that frequent proofreading consumes more GTP and slows translation [Gupta and Schlessinger, 1976, Thompson and Stone, 1977, Zengel et al., 1977, Ruusala et al., 1982, Bohman et al., 1984, Ehrenberg et al., 1986] (Ch. 1.6.1). Importantly, both of these processes relate the sequence of an mRNA to the duration of a translation cycle, providing a mechanism for how the ribosome modulates translation elongation rates.

Although ribosomal proofreading mechanisms clearly affect translation rates, it has also been suggested that mRNA secondary structure matters. Importantly, mRNA stability is sequence-dependent, and synonymous codons can give rise to different structures [Jou et al., 1971, Mita et al., 1988]. Predicting true *in vivo* secondary structures is difficult [Ramos and Laederach, 2014], however, so the extent to which they might stall ribosomes is unknown. Nevertheless, the ribosome has helicase activity [Takyar et al., 2005], which simultaneously suggests that the ability to unwind helical structures in mRNA is evolutionarily adaptive and that the ribosome is able to deal with at least some secondary structures effectively. Recent experiments have suggested that

the mRNA structure can affect translation rates [Yang et al., 2014, Firnberg et al., 2014], so it is possible that codons affect translation rates both through ribosomal mechanisms of protein synthesis as well as by modulating mRNA structure.

1.4 Translation is discontinuous

If translation initiation and elongation rates are both constant, then ribosomes will bind and process mRNA in defined intervals, producing a set spacing between the ribosomes on a given mRNA. Changes to the spacing between ribosomes can be detected by measuring the electrophoretic mobility of the polysome through a gel. It was shown that this mobility changes upon addition of translation inhibitors, indicating that the average polysome spacing of *E. coli* ribosomes can be artificially decreased or increased [Dahlberg et al., 1973]. Similarly, nascent proteins sampled from uniformly spaced ribosomes would equally populate all intermediate nascent protein lengths.

A pause site on the mRNA would, on average, spend more time bound to a ribosome than other sites. The lengths of nascent peptides collected from such a population would not be uniformly distributed. In extreme cases a ribosome pause causes a local “bottleneck” in which a down-

stream ribosome blocks the progression of upstream ribosomes [Wolin and Walter, 1988]. This situation causes a ribosome queue to form, changing the spacing between ribosomes even more. The mRNA upstream the pause site collects additional ribosomes, with reduced spacing, while the downstream mRNA has fewer ribosomes with greater spacing between them. These effects of discontinuous translation were succinctly explained by Itano using the metaphor of cars traversing a toll bridge [Itano, 1968]. However, the natural extent of ribosome queue formation was still unknown.

Preliminary evidence for a variable elongation rate *in vivo* was presented in the experiments which demonstrated that translation progresses from the 5′ to the 3′ end (N- to C-terminally) but the data were not quantitative [Dintzis, 1961]. It was subsequently shown that combined treatment with translation initiation and elongation inhibitors does not produce uniform β-galactosidase induction curves [Talkad et al., 1976]. This result was attributed to the existence of fast and slow ribosomes translating the *lacZ* message. Discontinuous translation was demonstrated in globin chains by quantifying the length distribution of nascent globin peptides using gel chromatography [Protzel and Morris, 1974]. Two competing explanations for translation rate differences, the formation of local mRNA structure which inhibits ribosome progression

and the depletion of tRNAs, were discussed. Arguments were made in favor of the role of mRNA secondary structure, not tRNA concentration [Chaney and Morris, 1978], and similar conclusions were drawn for bacteriophage MS2 coat protein [Chaney and Morris, 1979].

However, other experiments showed that tRNA concentrations can matter. Two groups separately demonstrated how the synthesis of silk worm fibroin in rabbit reticulocyte lysate fails to reach completion *in vitro*. Separation of nascent fibroin chains by SDS PAGE confirmed that some synthesis intermediates are more abundant than others, and that this effect is not due to selective degradation. These reactions cannot be rescued by the addition of dog brain tRNA, which has an "average" isoacceptor distribution, but *are* rescued by adding silk gland tRNA, which is adapted to fibroin synthesis [Lizardi et al., 1979, Chavancy et al., 1981].

The demonstration of reproducible pauses during the synthesis of silk worm fibroin, as well as their "rescue" following a change in the tRNA availability, strongly suggests that some codon/tRNA combinations are more prone to slow down the ribosome. However, this experiment has important shortcomings. Silk worm fibroin is translated in a highly specialized organ, in which fibroin makes up over 80% of all protein [Tashiro et al., 1968], whereas a typical cell could be more heterogeneous. Fur-

thermore, the tRNA content of these cells is highly specialized, with a small number of tRNAs accounting for the vast majority, and the fibroin amino acid sequence is itself highly repetitive, with four residue types accounting for 90% of the total [Fournier, 1979, Zhou et al., 2000]. Finally, the experiments were performed *in vitro*, providing greater access to manipulation but distancing the system from the physiological state. Thus, the same characteristics which made the system experimentally tractable also made the results less general.

Other experiments, performed in bacteria *in vivo*, demonstrated discontinuous protein synthesis while avoiding the aforementioned shortcomings. It was argued that overexpression of *E. coli* TolC, a minor outer membrane protein, leads to ribosome pausing at a single rare AGA codon [Misra and Reeves, 1985]. Likewise, Colicin A (ColA) protein, when overexpressed in either *C. freundii* or *E. coli*, was shown by SDS PAGE to produce protein synthesis intermediates to varying degrees [Varenne et al., 1982]. That these are translation intermediates, and not degradation products, was confirmed by the gradual disappearance of the lower molecular weight species following radioactive pulse-chase experiments. Other colicins showed slightly different patterns, as did *E. coli* pro-ompA and β-lactamase. Similarly, intermediates were observed during the synthesis of maltose-binding protein [Randall et al.,

1980].

The contributions of tRNA availability to ribosome pausing were tested by comparing the banding pattern with or without a hundredfold stimulation of colicin expression. High colicin expression would be expected to exacerbate pauses at codons recognized by low abundance tRNA, yet it was shown that the banding pattern persists, and was concluded that discontinuous colicin synthesis is not due to tRNA concentration perturbations [Varenne et al., 1984]. The lack of correspondence between predicted hairpin locations and the observed banding patterns also suggested that mRNA secondary structures cannot explain the observed pauses. Rather, the pauses during protein synthesis were ascribed to a stochastic search, during which the probability of the matching tRNA arriving at the A-site is determined by a competition with all the other tRNA molecules [Varenne and Lazdunski, 1986]. According to this model, adding increasing amounts of all tRNA would not actually alter their relative abundances, and therefore have no effect on the stochastic search.

This model was experimentally tested by inserting slow AGG codons into the chloramphenicol acetyltransferase (*cat*) gene [Robinson et al., 1984, Varenne et al., 1989]. In the latter study, four or six total AGG codons were inserted into the *cat* gene two different ways. One gene

contained four clustered AGG codons, while the other contained six AGG codons which were not clustered. The results demonstrated that there were no differences in protein expression levels when transcription levels of the *cat* mRNA were low, but that the production of Cat protein from the clustered allele underwent a dramatic decrease at high transcription levels. This effect was attributed to sequestration of the $tRNA^{Arg}$ within the ribosome, causing a depletion [Varenne et al., 1989], and suggesting that codon adjacency or context effects also affect ribosome pausing.

Other experiments on the *cat* gene were performed with *E. coli* cell free extracts. Many nascent proteins of intermediate length were visible following translation of the wild-type mRNA. As expected, the re-coding of a 16 codon sequence selectively removed a pause site from the mRNA, which manifested itself as the disappearance of a single band on an SDS polyacrylamide gel [Komar et al., 1999]. The quantification of nascent proteins has clearly shown that translation rates are not necessarily uniform.

1.5 Experimental methods to measure bacterial translation elongation rates

Measuring translation rates *in vivo* has been achieved with several methods, with trade-offs striking a balance between generality and precision. *E. coli* is the organism with the most intricately studied elongation rates, both in terms of the number of published data as well as mechanistic insight. Here, I describe these methods and summarize the measured translation rates (Table 1.1, p. 51).

The first estimates of the polypeptide chain growth rate were done with bulk cell assays. The amount of protein mass incorporated over time, multiplied by the fraction of active ribosomes (*e.g.* 0.77), yields an estimate for how much protein is incorporated per ribosome per unit time [Maaløe and Kjeldgaard, 1966]. Normalizing by the mass of an average amino acid (120 Da) then gives how many amino acids are incorporated per ribosome, and normalizing by the growth rate (μ) takes into account how total protein levels rise in proportion to cellular growth (Eq. 1.1, p. 25). Subsequent studies used similar reasoning to estimate the elongation rate of *E. coli* ribosomes synthesizing ribosomal

proteins [Schleif, 1967, Dennis and Bremer, 1974] or all proteins [Forchhammer and Lindahl, 1971]. The result of this analysis is susceptible to multiple sources of measurement error: 1. the molecular weight of the average amino acid, 2. the number of ribosomes, which required estimates of the rRNA mass per ribosome, the relative amounts of tRNA and rRNA in the cell and the assumption that mRNA is negligible, 3. a correction factor (∼75% – 85%) for inactive ribosomes and 4. the growth rate of the cells. Furthermore, the result is a population average of all nascent proteins. Nevertheless, this method consistently produced values between 10 and 20 a.a. residues per ribosome per second.

$$\text{a.a.}/(\text{active ribosome} \cdot \text{s}) = \frac{\frac{\text{protein}}{120} \times \mu \times \ln 2}{\frac{\text{rRNA}}{1.6 \times 10^6} \times 0.77} \tag{1.1}$$

Methods which avoided these issues were developed, in particular those which measure the synthesis rate of a single protein, β-galactosidase (encoded by gene *lacZ*). An average *E. coli* protein of 300 residues would be translated in only 15 – 30 seconds, making measurements extremely difficult, but β-galactosidase is 1,024 a.a. long. Also, the protein catalyzes an enzymatic reaction which can be measured with a spectrophotometer, and its synthesis is inducible by IPTG (isopropyl β-D-1-thiogalactopyranoside), allowing for precise timing. In a pioneering study, 5-methyl tryptophan was used to temporarily poison protein synthesis

without interrupting transcription [Kepes and Beguin, 1966]. After IPTG induction, *lacZ* messages reached their new steady state levels in a few minutes, but protein synthesis did not occur. Relief of the poison by the addition of large amounts of regular tryptophan restored ribosome activity. The time between the addition of tryptophan and the appearance of enzymatic activity, known as the delay time, is therefore the minimal time to synthesize a complete β-galactosidase tetramer. This delay time was linearly extrapolated, and it was found that between 80 and 90 seconds elapsed before new enzymatic activity was detectable. At the time of the study, the number of residues per β-galactosidase chain was unknown, but today we know that the chains are 1,024 residues long, and so can infer that the translation rate was between $\frac{1024 \text{ a.a.}}{90 \text{ s}} = 11.38 \frac{\text{a.a.}}{\text{s}}$ and $\frac{1024 \text{ a.a.}}{80 \text{ s}} = 12.80 \frac{\text{a.a.}}{\text{s}}$.

A second method uses radioactive pulse-chase labeling of nascent protein. The experiment begins by adding a pulse of radioactive amino acid, which is diluted with unlabeled amino acids after a set amount of time. Any increase in radioactivity over time is due to labeled amino acids which diffused into cells, were ligated to tRNA, reached a ribosome and were incorporated into nascent polypeptide. Periodic sampling from the culture, following the chase, shows that a maximal amount of radioactivity is reached when the last protein containing radioactive

amino acids is completed. The time when maximum radioactivity is observed corresponds to the time necessary to synthesize a protein. Early experiments of total ribosome samples roughly estimated that the average *E. coli* protein took 5 s to synthesize [McQuillen et al., 1959].

Of course, this time depends on the length of the protein, and so better resolution demanded either analyzing purified protein, or separating proteins by length. Lacroute and Stent induced expression of β-galactosidase, performed a pulse-chase experiment, and then purified the enzyme. They estimated a delay time of 80s, and thus a translation rate of 12.8 a.a. s^{-1} [Lacroute and Stent, 1968]. Gausing modified a pulse-chase method [Bremer and Yuan, 1968] in which the samples are run on a denaturing gel and separated by molecular weight [Gausing, 1972]. These data produce averages of similarly sized proteins, and gave similar estimates of 11.4 a.a. s^{-1}. Further refinements of this last method separate proteins on 2D gels [Pedersen, 1984] or with more sophisticated data analyses [Sørensen and Pedersen, 1991, Sørensen and Pedersen, 1998], yielding an estimate of 12.5 a.a. s^{-1} [Sørensen et al., 1989].

A third method to measure the translation rate is a variation of the β-galactosidase delay time assay. Unlike Kepes' experiments, no initial protein synthesis inhibition is necessary. It was reasoned that, af-

ter inducer addition, the protein level increases quadratically [Schleif et al., 1973]. Thus additional time-points measured beyond the first appearance of additional enzymatic activity can be used to fit the square root of enzymatic activity versus time. The authors estimated a delay time of 1.2 min, which corresponds to a translation rate of 14.22 a.a. s^{-1}. They also performed the measurement as developed by Gausing and observed similar rates of 10–13 a.a. s^{-1}, showing good agreement between these two methods. However, it has been argued that this method can give erroneous results when transcription elongation is slowed down [Sørensen et al., 1994], or when the time for the newly-synthesized proteins to fold is slow [Kepes, 1963].

1.6 Parameters which affect translation elongation rates

The above methods have been used to probe many aspects of how translation rates are regulated. These include the effect of the medium and the cellular growth rate [Engbaek et al., 1973, Dennis and Bremer, 1974, Young and Bremer, 1976, Pedersen, 1984], infection by bacteriophage T4 [Gausing, 1972], temperature [Goldstein et al., 1964, Farewell

and Neidhardt, 1998], polyamine concentration [Coffman et al., 1971, Morris and Hansen, 1973, Jorstad and Morris, 1974], tRNA modifications [Krüger et al., 1998], ribosomal accuracy and streptomycin sensitivity [Galas and Branscomb, 1976, Gupta and Schlessinger, 1976, Zengel et al., 1977, Andersson et al., 1982, Bohman et al., 1984, Ruusala et al., 1984, Tubulekas and Hughes, 1993], and mRNA sequence [Pedersen, 1984, Sørensen et al., 1989, Sørensen and Pedersen, 1991, Spencer et al., 2012]. These last two properties are the most relevant to this work and will be developed in more detail.

1.6.1 Accuracy and streptomycin resistance

Kinetic proofreading explains how template-driven polymerization can maintain fidelity beyond that which would be explained by thermodynamics alone [Hopfield, 1974, Ninio, 1975]. It was predicted that changes in the translation elongation rate would be met with corresponding changes in translation accuracy [Ninio, 1974], which is true for some ribosomal mutants. The antibiotic streptomycin has been key to understanding ribosomal fidelity. Streptomycin-resistant bacteria (Sm^R) were found to have mutations to ribosomal protein S12 and are designated *strA*. These bacteria can grow in the presence of the antibiotic, but also suppress the reading of stop codons [Gorini and Kataja, 1964].

The ribosomal screen mechanism was first proposed, suggesting that the ribosome can bind tRNA while simultaneously modulating codon-anticodon interactions, such that in the case of mismatches the tRNA will be discarded [Gorini, 1971].

E. coli cells resistant to streptomycin were tested for their translation elongation rates. It was shown, using the β-galactosidase delay time assay, that resistant bacteria have slower ribosomes [Galas and Branscomb, 1976, Andersson et al., 1982], while exposure to the drug increased translation rates in a concentration-dependent manner [Gupta and Schlessinger, 1976]. These results were confirmed using the β-galactosidase assay as well as the radioactive pulse-chase method [Zengel et al., 1977]. Sub-classes of streptomycin-resistant strains were further explored. The translation rates of streptomycin-dependent (Sm^D), pseudo-dependent (Sm^P) and resistant (Sm^R) bacteria were compared using the β-galactosidase method, and it was clear that these mutations to the ribosome altered the translation rate, which in turn was affected by the presence or absence of the drug [Bohman et al., 1984, Ruusala et al., 1984].

Other mutations besides those to the ribosome can alter both ribosomal accuracy and rate. A study compared the complementary effects of mutations to S12 and to EF-Tu and found that single mutations to either

protein reduced both the elongation rate and growth rate [Tubulekas and Hughes, 1993]. However, various combinations of the mutations were largely able to restore growth and translation. They also confirmed that a class of non-restrictive mutants have unperturbed elongation rates, confirming that some, but not all, changes to accuracy have an effect on the translation rate.

1.6.2 Codon composition

A series of careful studies during the 1980s and early 1990s clearly demonstrated that codon composition has an effect on translation rates. The comparison of five *E. coli* genes, two with "rare" codons and three with "frequent" codons, was performed using pulse-chase analysis. It was found that mRNAs with rare codons are translated more slowly [Pedersen, 1984]. This effect was exacerbated by high expression levels, possibly indicating the depletion of less abundant tRNA. In all cases, translation rates increased at higher cellular growth rates. Followup experiments, using inserts of different codons within the *lacZ* gene, revealed a threefold difference in the translation rates of slow and fast codons [Sørensen et al., 1989]. Interestingly, no effect of mRNA secondary structure was observed.

These experiments culminated with extremely detailed analyses of

the translation rates of a few individual codons [Sørensen and Pedersen, 1991]. As before, different inserts were placed within the *lacZ* gene; however, a few clever tricks allowed for increased sensitivity of the analyses. The inserts produced a shift in the migration of the completed β-galactosidase protein, which allowed for the simultaneous comparison of mutated and wild-type mRNA translation rates within a same culture, controlling for loading error. The researchers were able to decompose the relative contributions of individual types of codon within the inserts, resulting in absolute translation rates for codons GAA, GAG, CCG and CGA. GAA and GAG both encode glutamate, differing only at the wobble position. Crucially, they share the same tRNA, and so effects of tRNA concentration are controlled for. This experiment provides good evidence that wobble translation is slow, as the GAA codon is translated with an average rate of 21.6 a.a. s^{-1}, whereas GAG is translated at 6.4 a.a. s^{-1}.

More recently, the same technique was used to assess the contributions of codons within the Luc mRNA. Different synonymous codon repertoires were employed: a "fast" Luc mRNA, using both cognate and high copy tRNA, the wild-type (WT) Firefly sequence, and a "CBF" sequence with the most frequently occurring codons from the *E. coli* genome [Spencer et al., 2012]. As expected, the translation rate de-

pended on the mRNA sequence. A “slow” mRNA was also tested, but low expression precluded a successful translation rate measurement.

Several experiments which did not provide absolute translation rate measurements deserve a special mention. A method was developed to monitor the coupling between transcription and translation within the *pyrE* attenuator system [Bonekamp et al., 1985]. Different codons can be inserted into a *lacZ* reporter system under control of the system, and the degree of expression indicates whether or not the translation elongation rate changes. The method revealed that AGG codons are slow, regardless of the gene expression level [Bonekamp and Jensen, 1988], a result which is consistent with the observation that AGG codons are extremely rare and can cause frameshifting [Spanjaard and van Duin, 1988]. Finally, it was also shown that translation rates depend on codon composition; however, the relationship is not as straightforward as might be expected [Bonekamp et al., 1989]. This result should not be interpreted to mean that tRNA levels or codon composition have no effect, but rather that there are also other contributors to translation rates.

1.7 Is co-translational protein folding kinetically regulated?

After the discovery that proteins are linearly synthesized, from the N-terminus to the C-terminus [Dintzis, 1961], and experiments which suggested that *E. coli* β-galactosidase chains can be found in a tetrameric form while still ribosome-bound [Zipser and Perrin, 1963, Kiho and Rich, 1964], it was proposed that protein folding pathways could depend on proximity to the ribosome [Levinthal, 1968]. There has since been much evidence of co-translational folding (*e.g.* [Fedorov et al., 1992, Frydman et al., 1999, Nicola et al., 1999, Holtkamp et al., 2015], reviewed by Fedorov and Baldwin [Fedorov and Baldwin, 1997]), and the ribosome itself has been implicated in promoting the folding of nascent or completed proteins [Kudlicki et al., 1997, Kaiser et al., 2011]. However, both the degree and extent to which this process is biologically controlled or beneficial remain unclear. Numerous attempts have been made to computationally search for synonymous codons which correspond to specific protein secondary structures, yet no unifying principles have emerged [Adzhubei et al., 1996, Brunak and Engelbrecht, 1996, Thanaraj and Argos, 1996a, Thanaraj and Argos, 1996b, Xie and Ding, 1998, Gupta et al., 2000, Saunders and Deane, 2010].

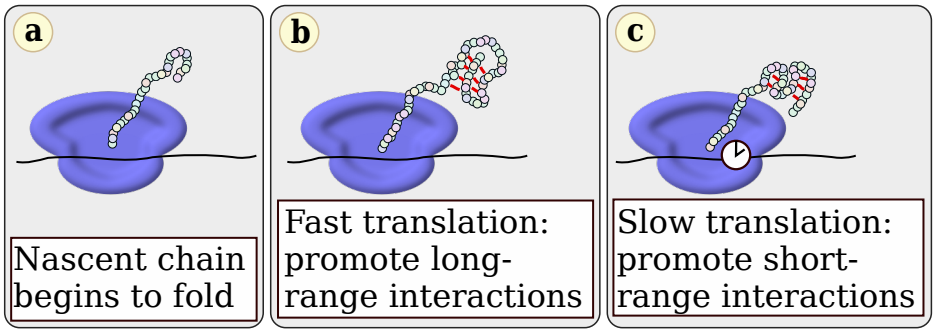


Figure 1.3: The translation rate might alter the distance-dependent formation of contacts within a nascent polypeptide. The emergence of a nascent polypeptide from the ribosome tunnel could be as fast as 20 residues per second or as slow as 5 residues per second under identical cellular conditions. a. A nascent protein begins to adopt secondary structure as it emerges from the ribosome. b. Fast translation quickly grows the nascent chain, promoting the formation of contacts between residues far apart in the primary sequence. c. A translation pause allows extra time for the formation of more local contacts, before additional amino acids are incorporated.

Nevertheless, the effects of mRNA sequence on co-translational protein folding have been experimentally tested (reviewed by [Marin, 2008, Tsai et al., 2008, Zhang and Qian, 2011, Sauna and Kimchi-Sarfaty, 2011, Deane and Saunders, 2011, Spencer and Barral, 2012, Chaney and Clark, 2015]). However, the majority of these studies have only indirectly demonstrated such a connection [Crombie et al., 1992, Komar et al., 1999, Cortazzo et al., 2002, Evans et al., 2008, Kimchi-Sarfaty et al., 2007, Pechmann et al., 2014]. It has been particularly challenging to simultaneously and quantitatively show that the folding of the protein is impaired, that the translation rate has been altered and that the rate is genetically encoded.

Several studies have more convincingly shown such a relationship. In one case, codons which were predicted to be either fast or slow made *E. coli* SufI protein more or less sensitive to proteolysis by proteinase K [Zhang et al., 2009]. Another study looked at the effects of coding on the translation of a hybrid protein, in which the N- and C-terminal domains compete for folding with a shared middle domain. It was shown that "rare" codons immediately upstream of the C-terminal domain promote the folding of the middle domain with the N-terminal domain [Sander et al., 2014]. However, in both examples the absolute translation rate differences were small or undetectable.

Recently, a circadian role for protein folding regulated by slow codons has emerged. This role has been demonstrated in both bacteria and eukaryotes, even though their proteins are not conserved [Xu et al., 2013,Zhou et al., 2013,Fu et al., 2016]. Replacing wild-type codons with faster ones in the *Synechococcus elongate kaiBC* mRNA eliminated a temperature-sensitive regulatory response [Xu et al., 2013], suggesting that the slow codons are adaptive to low temperature, while removing slow codons in the *Neurospora FRQ* gene ablated the maintenance of a clock [Zhou et al., 2013]. Interestingly, in *Neurospora* these "slow" codons naturally occur within the part of the *FRQ* mRNA which encodes intrinsically disordered regions of the FRQ protein, an observation whose significance is unclear. Other results have been shown for bovine lens protein γB-crystallin, whose mRNA was suggested to have an evolved codon pattern which optimizes protein folding [Komar and Jaenicke, 1995]. Proteinase K sensitivity data, as well as NMR data, suggest that synonymous codons can improve or worsen the protein's co-translational folding [Buhr et al., 2016].

Firefly Luciferase (Luc) is also a popular model protein for protein folding. In two studies, one in *Neurospora* and the other in *E. coli*, it was suggested that Luc protein folding is improved after reducing the translation rate by synonymously re-coding the mRNA [Yu et al.,

2015, Spencer et al., 2012]. The translation rates in the *Neurospora* cell lysates were between 2 and 3 a.a. per sec [Yu et al., 2015], while in *E. coli* they spanned roughly 10 a.a. per sec to 20 a.a. per sec [Spencer et al., 2012]. In both examples, Luc specific activity differences were observed depending on which codons were present in the mRNA sequence, and this result was interpreted to mean that the folding of the protein was impaired due to changes in translation rates.

Currently, many observations linking translation rates to protein folding have been made, but with no unifying principle which predicts which residues, or which translation rates, are affected. This issue might be remedied with more information about protein structures and their translation rates, yet both are difficult to obtain. However, differences in codon-specific translation rates might now be detectable using high-throughput sequencing methods.

1.8 Using ribosome profiling to understand translation pauses

The translation rate has been measured for individual codons, genes and transcriptomes, yet measuring the specific translation rates of all

61 sense codons has been a longstanding challenge. Ribosome profiling, a high-throughput sequencing technique [Ingolia et al., 2009], is now being used to address this question. In this experiment, mRNA bound by ribosomes is extracted and subjected to nuclease digestion. Only the stretches of mRNA residing within ribosomes are protected from the nuclease. Many such fragments are then sequenced, and the frequencies of identical sequences within the dataset reveal the relative translation levels of the corresponding mRNAs, thereby quantifying their relative translation levels. Ribosomes spend more time translating slow codons, so it follows that more sequencing reads of a specific locus reflect slower translation. Two of the first ribosome profiling experiments clearly showed that the data produce regions with high or low ribosome occupancy [Ingolia et al., 2009, Oh et al., 2011]. In practice, however, deriving relative translation rate differences using ribosome profiling has been difficult [Gritsenko et al., 2015, Wang et al., 2016b, Weinberg et al., 2016, Liu and Song, 2016], and synonymous changes to an mRNA were only recently shown to change the ribosome profile [Yu et al., 2015]. Here is a brief review of the ribosome profiling literature, with an emphasis on understanding how codons regulate translation elongation.

1.8.1 Meta analyses of codon occupancy

Analysis of all footprints, averaged together, is known as meta analysis. Such aggregated data can be used to look for major trends which might not be otherwise apparent. This approach is valid when comparing codon-specific translation rates as long as the "rules" of translation are the same between transcripts. One way to evaluate ribosome occupancy of the codons is to count them relative to some expected value. However, this procedure is not trivial for two reasons. First, a typical ribosome footprint is ∼10 codons (28 nt) long [Steitz, 1969, Wolin and Walter, 1988], but only three of these codons are located within the ribosomal active sites at any given time. Therefore, to understand differences in codon translation rates it is crucial to correctly assign the A, P and E sites within each footprint. Second, determining the reference codon frequencies is likewise important. These can be derived by comparing two sets of experiments (*e.g.* with or without treatment), or from the codon frequencies of the genome, weighted by expression level. Numerous studies of this type have been performed, in both bacteria and metazoans, each revealing different aspects of translation rate variation.

As discussed above, translation of the 61 sense codons differs in many ways, but most striking is that no organism has 61 matching tRNAs

[Ninio, 1971, Chan and Lowe, 2009]. Since codons lacking a perfectly matching, WC tRNA can still be translated, they must be translated by tRNAs which are imperfect matches. Imperfect matches at the third codon position are called wobble [Crick, 1966], and it has been demonstrated that wobble recognition is slow [Kato et al., 1990, Krüger et al., 1998, Sørensen and Pedersen, 1991, Spencer et al., 2012]. To test whether all wobble codons are slow, ribosome profiling experiments compared the incidence of different codons within the ribosome A site, in both *C. elegans* and HeLa cells [Stadler and Fire, 2011]. The P site in *C. elegans* was assigned to a location within the ribosome by analyzing initiator tRNA occupancy; this position was offset by one codon in HeLa ribosomes. It was found that wobble codons are enriched, confirming that wobble translation is slow and that ribosome profiling can detect such differences. Similar results were replicated in yeast [Lareau et al., 2014, Wang et al., 2016a].

The first profiling studies of *E. coli* did not show a role for wobble translation [Li et al., 2012]. In fact, all codons had nearly equal ribosome occupancy (except for serine, see below), and the authors concluded that codon translation rates in *E. coli* are not significantly different. This conclusion was problematic for two reasons. The first is that codon frequencies are not uniform. If all codon translation rates

are the same, then the observed footprint codon frequencies should be proportional to the transcriptome frequencies, yet their data show the opposite. Second, their observations were reconciled with an A site assignment which is different from eukaryotes', and is relative to the 3′ end of the footprint [Martens et al., 2015]. When this property was taken into account, the codon frequencies closely matched the expected values (data not shown). Still, slow wobble translation in *E. coli* has not yet been conclusively demonstrated using ribosome profiling.

Before the recognition step takes place within the ribosome, tRNAs must first be activated. Functional tRNAs are charged with an amino acid, and aminoacylation levels depend on the metabolic state of the cell. Starvation conditions lead to depletion of translation-competent tRNA, with implications for the translation rate [Elf et al., 2003, Dittmar et al., 2006], whereas supplementation of amino acids is expected to produce optimal translation and generally good growth conditions. Consistent with this idea, selective depletion of specific amino acids led to ribosome pausing at those codons [Subramaniam et al., 2014]. It was therefore surprising that *E. coli* grown in LB (with abundant amino acids) showed signs of serine limitation [Li et al., 2012]. Under these conditions, serine codons are over-represented in the ribosomal A site, consistent with slow translation [Martens et al., 2015]. This effect was ascribed to pref-

erential serine metabolism as a carbon source, which depletes cellular serine levels. Another explanation is that active serine detoxification keeps serine levels perpetually low [Avcilar-Kucukgoze et al., 2016]. The absolute change in the translation rate of serine codons under such conditions is still unknown, and whether serine codons act as conditional regulators of the translation rate remains unexplored. Nevertheless, these results demonstrate that ribosome profiling has revealed strong, unanticipated codon-dependent ribosome pausing.

Other codons are expected to give rise to pauses because of the corresponding amino acid. Proline, a non-standard "imino" acid, behaves differently from other amino acids when incorporated into polypeptides. *In vitro* experiments had shown kinetic impairment to the translation of proline codons due to geometric effects, and these were confirmed using ribosome profiling in yeast [Artieri and Fraser, 2014a, Gardin et al., 2014, Martens et al., 2015]. Ribosome profiling of wild-type *E. coli* did not reveal any general effect [Martens et al., 2015] (Ch. 2, p. 52), but the deletion of a special elongation factor for prolines (EF-P) also resulted in proline enrichment in the footprints [Woolstenhulme et al., 2015]. These results show that *E. coli* proline codons are not slow under normal circumstances, but proline incorporation in *S. cerevisiae* is slower than average.

The translation rate is often attributed to the kinetics of amino acid incorporation. However, forces outside the ribosomal catalytic site have been suggested to slow down translocation. Bacterial ribosomes recognize start codons with the help of the Shine-Dalgarno sequence, whose consensus sequence is 5′-AGGAGG-3′. This sequence binds an rRNA bearing the sequence 3′-UCCUCC-5′. Although the binding of the ribosome to the mRNA takes place during initiation, it is conceivable that the rRNA could also bind the mRNA at subsequent downstream positions. Enrichment of high GC content sequences in ribosome profiling data was presented as evidence of this mechanism [Li et al., 2012, Schrader et al., 2014]. However, this signal is likely an artifact of library preparation [Martens et al., 2015, Mohammad et al., 2016] and has not been replicated in-vitro [Borg and Ehrenberg, 2015].

Charged residues were also proposed to stall ribosomes. Comparisons between ribosome densities, in *S. cerevisiae* footprints, and the locations of charged residues showed statistically significant relationships [Charneski and Hurst, 2013]. The magnitude of the effect increased when multiple positively charged residues were near one another. Since it has been estimated that the ribosome tunnel is negatively charged, the clustering of positively charged residues would bind the tunnel and present an energetic barrier for translocation. This result

is intriguing because it also suggests a mechanism for the eukaryotic poly-A tail (encoding positively charged lysine) in halting frameshifted ribosomes. In fact, it was recently shown in yeast that ribosome stalling at poly-lysine stretches induces mRNA cleavage [Guydosh and Green, 2017].

A convincing experiment compared ribosome pausing before and after a treatment which specifically targets the translation of histidine codons. Histidine biosynthesis is inhibited upon treatment with 3-amino-triazole, leading to a depletion of the charged $tRNA^{His}$ [Klopotowski and Wiater, 1965]. Much as serine depletion causes a rise in serine codons in ribosome footprints in *E. coli*, this depletion in *S. cerevisiae* caused an increase in histidine codons [Guydosh and Green, 2014, Gardin et al., 2014, Lareau et al., 2014]. However, while the *E. coli* serine signal was located to a specific location within the footprints, this yeast histidine signal was spread throughout [Martens et al., 2015]. This experiment produces a clear comparison of the perturbation of a single group of codons and nicely demonstrates the expected enrichment.

In summary, the detection of ribosome pausing at specific types of codons, using ribosome profiling, is well established. In the future, an investigation into whether these general effects occur preferentially in certain genes, or certain mRNA positions, would be highly informative.

For example, it might be that some codon-dependent ribosome pauses are contextually modulated [Gamble et al., 2016].

1.8.2 Meta analyses of ribosome density along transcripts

Meta gene analysis is the assembly of an aggregated “gene” whose footprints are summed along the coding region, beginning at the start codon. Unlike meta footprint analysis, this method keeps track of the exact position where the ribosome was on the transcript, and so captures average patterns of ribosome distributions along the mRNAs. In two of the first ribosome profiling studies it was shown that metazoan footprints follow a three base-pair periodicity consistent with the translocation of the ribosome [Ingolia et al., 2009, Stadler and Fire, 2011]. This signal, absent from mRNA-Seq data, is strong evidence that ribosome profiling measures translation.

Meta gene analysis has also been used to study changes in ribosome densities along the mRNA. Translation elongation inhibitors produce an increase in ribosome density upstream of the start codon, allowing mRNAs to load a ribosome but preventing elongation [Ingolia et al., 2009]. It was then shown that this 5′ ribosome density can propagate

along transcripts [Ingolia et al., 2011]. To measure the average ribosome movement, a combination of translation initiation and elongation inhibitors was employed. In this experiment, a translation initiation inhibitor (harringtonine) is added at time zero, and after a variable delay is followed by an elongation inhibitor (cycloheximide). During the delay, ribosomes will gradually become depleted near, and downstream of, the start codon as they continue to translocate. When the data for all transcripts were combined, the growing ribosome density gap became apparent. These data were used to calculate the average translation rate in *S. cerevisiae*, and the resulting value of 5.6 codons s^{-1} is in good agreement with estimates from mammalian cell lines [Boström et al., 1986]. Curiously, regardless of which genes were analyzed, no relationship was found between the codon content of the mRNA and the propagation rate of the ribosome density. This result either suggests that the codon composition of genes does not influence translation elongation rates in *S. cerevisiae*, at odds with previous findings, or that ribosome profiling is unable to capture gene-specific differences.

The first ribosome profiling studies only analyzed footprints ∼28 nt long, as this length had been experimentally demonstrated [Steitz, 1969, Wolin and Walter, 1988]. Based on early work on mRNA protection within ribosomes, ribosome footprints ∼28 nt long were preferentially

selected during library preparation. However, selection of smaller mRNA fragments revealed that ribosome footprints from *S. cerevisiae* comprise two different length distributions [Lareau et al., 2014]. These two classes reflect different conformations of actively translating ribosomes and can be selectively amplified. One conformation (~28 nt footprints) is favored following treatment with cycloheximide, while the other conformation (~22 nt footprints) is favored by anisomycin. Separating these two classes of footprints into their own meta-gene analysis reveals that both exhibit 3 bp periodicity.

Translation inhibitors are an important part of many ribosome profiling protocols because they ensure that ribosomes cease activity during preparation steps. Given that translation rates are rapid, even a one-second delay between cell harvesting and translation arrest could lead to loss of precision and erode the ability to ascertain single-codon resolution. For this reason treatment with liquid nitrogen (LN2) has also been preferred, as it acts quickly and halts all chemical processes equally. In fact, recent work has shown that any treatment with translation inhibitors prior to harvesting causes uneven ribosomal arrest and degrades the local signal [Hussmann et al., 2015]. In this paper the authors analyzed many profiling studies and categorized them by their use of cycloheximide. They find that codon occupancy in the absence of the

inhibitor matches expectations from the tRNA adaptation index (tAI) far better than with inhibitor. They also demonstrate how different cycloheximide treatment regimes lead to propagating “waves” of ribosome density, consistent with residual ribosome activity even in the presence of the inhibitor.

1.9 Perspectives

Translation’s central role in biology has made it a vital area of research, but its sheer complexity leaves us with some basic questions unanswered. We still don’t know the translation rates of most codons, or how these rates change depending on conditions. The data thus far suggest that some codons have much higher translation rates than others, but until this issue is resolved it will be difficult to know to what extent the translation rate itself is used as a biological mechanism. Ribosome profiling might finally bring the necessary details that were, until now, out of reach. This knowledge could be definitive in addressing the extent to which the mRNA encodes not only the amino acid sequence but other information as well.

Table 1.1: Experimentally determined bacterial translation rates

Rate (a.a. s^{-1})	Comment	Reference
11.3		[Kepes, 1963]
0.14 – 0.24	0°C	[Goldstein et al., 1964]
14 – 17	*S. typhimurium*	[Maaløe and Kjeldgaard, 1966]
11.4 – 12.8		[Kepes and Beguin, 1966]
11 – 15		[Schleif, 1967]
12.8		[Lacroute and Stent, 1968]
10.2 – 11.4		[Coffman et al., 1971]
12 – 21		[Forchhammer and Lindahl, 1971]
11.4, 5.6	WT, T4 infection	[Gausing, 1972]
14.2, 10 – 13		[Schleif et al., 1973]
16 – 20		[Engbaek et al., 1973]
17, 10	WT, - polyamine	[Morris and Hansen, 1973]
17.5, 9.5	WT, - polyamine	[Jorstad and Morris, 1974]
11 – 20		[Dennis and Bremer, 1974]
10.9 – 14.2		[Dalbow and Young, 1975]
12.8, 9.2	WT and Str^R	[Galas and Branscomb, 1976]
8.5, 3.1 – 4.9	WT, + streptomycin	[Gupta and Schlessinger, 1976]
12 – 17		[Young and Bremer, 1976]
6.6 – 7.42, 2.6 – 7.3	WT, + streptomycin	[Zengel et al., 1977]
17 – 19		[Andersson et al., 1982]
15, 5 – 11, 10	WT, Sm^P, Sm^D	[Ruusala et al., 1984]
7 – 17		[Pedersen, 1984]
7.8 – 14	Various streptomycin	[Bohman et al., 1984]

12.5		[Sørensen et al., 1989]
21.6, 6.4, 5.8, 4.2	GAA, GAG, CCG, CGA	[Sørensen and Pedersen, 1991]
7.4 – 12.9	WT, ribosome mutants	[Tubulekas and Hughes, 1993]
1.9 – 47	$tRNA^{Glu}$ mutants	[Krüger et al., 1998]
7.5, 12.9, 16.9	28°C, 37°C, 42°C	[Farewell and Neidhardt, 1998]
9.8 – 19.2	Re-coded Luciferase	[Spencer et al., 2012]
12.5		Ch. 3, p. 108

Chapter 2

Ribosome A and P sites revealed by length analysis of ribosome profiling data

The high-throughput sequencing of nuclease- protected mRNA fragments bound to ribosomes, a technique known as ribosome profiling, quantifies the relative frequencies with which different regions of transcripts are translated. This technique has revealed novel translation initiation sites with unprecedented scope and has furthered investigations into the connections between codon biases and translation rates. Yet the location of the codon being decoded in ribosome footprints is still unknown, and has been complicated by the recent observation of

footprints with non-canonical lengths. Here we show how taking into account the variations in ribosome footprint lengths can reveal the ribosome aminoacyl (A) and peptidyl (P) site locations. These location assignments are in agreement with the proposed mechanisms for various ribosome pauses and further enhance the resolution of the profiling data. We also show that GC-rich motifs at the 5′ ends of footprints are found in yeast, calling into question the anti-Shine-Dalgarno effect's role in ribosome pausing.

2.1 Introduction

Protein production, one of the most fundamental biological processes, determines the ability of cells to adapt to their environments. Variations in gene expression allow a cell to modulate which proteins are produced, which in turn establishes the levels of key cellular actors. For example, bacteria have evolved intricately regulated protein synthesis and are able to complete the production of new protein within minutes of stimulation [Schleif et al., 1973]. Of the three stages of translation (initiation, elongation and termination), initiation rates are the primary determinants of protein abundance [Mathews et al., 2007, Schwanhäusser et al., 2011] and therefore have the largest impact on protein

production. There is increasing evidence that differential translation "elongation" influences the folding of nascent polypeptides [Siller et al., 2010, Spencer et al., 2012, Marin, 2008], but the extent to which translation elongation is regulated, and how this regulation affects protein folding, is poorly understood.

Pioneering studies using the ribosome profiling technique have addressed key issues of translation by quantitatively sequencing nuclease-protected, ribosome-bound transcript fragments [Ingolia et al., 2009, Li et al., 2012, Ingolia et al., 2011]. The patterns underlying these ribosome 'footprints' give clues to nature of the forces behind translation elongation and the relationships between translation kinetics, cellular milieu and the genetic code. The canonical ribosome footprint was determined to be 28 nt long [Steitz, 1969, Wolin and Walter, 1988]. As revealed by deep sequencing, however, there are small but substantial populations of footprints which are shorter [Lareau et al., 2014]. *A priori*, there is no clear way to compare footprints of different lengths and such differences obscure the exact position of a ribosome along a message. The extra nucleotides could conceivably occur at either end of the ribosome, an issue that remains unresolved.

Although ribosome profiling data have proven to be highly informative, the location of the codon being decoded within a footprint is still

unclear. A growing body of evidence suggests that, under certain conditions, the translation of specific amino acids can impair ribosomal processivity, including proline [Pavlov et al., 2009, Artieri and Fraser, 2014a], serine [Li et al., 2012] and histidine [Guydosh and Green, 2014, Lareau et al., 2014]. Prolines are thought to impair ribosome processivity during peptidyl transfer, whereas pausing at serine and histidine codons is the result of an increased delay while waiting for ternary complex at the A site. Yet other ribosome pausing, due to the anti-Shine-Dalgarno (aSD) effect [Li et al., 2012, O'Connor et al., 2013], can presumably occur because of interactions outside the ribosome active site region and thereby require other mechanisms. The relative contributions of all these effects to ribosome pausing are unclear. We hypothesized that a better understanding of how the sequence of a footprint, which reports on the codons being translated, is related to the number of times it occurs in the sequencing library and therefore the pausing propensity, could give us insight into both the location of the ribosome A site and the determinants of ribosome pauses. To test this hypothesis, we analyzed published datasets, both from yeast and *Escherichia coli*.

2.2 Materials And Methods

2.2.1 Computing codon frequencies

Briefly, using previously reported datasets, ribosome footprints were processed by determining where they align to the genome with Bowtie [Langmead et al., 2009] and then categorized by length (in codons) using a Perl program. For each length category, the frequencies of codons at each position within the footprint were tabulated, where the position is defined relative to either the 5′ or 3′ end of the footprint. The resulting were structured as 64 2D matrices, where each matrix represents one of the 64 codons and each matrix position represents a codon of a particular footprint length and at a particular footprint position (Fig. 2.1, p. 57).

2.2.2 Computational processing of ribosome profiling sequencing libraries

Data obtained from *E. coli* cells grown in LB medium are from SRR364368 and SRR364370 (Gene Expression Omnibus (GEO) accession number

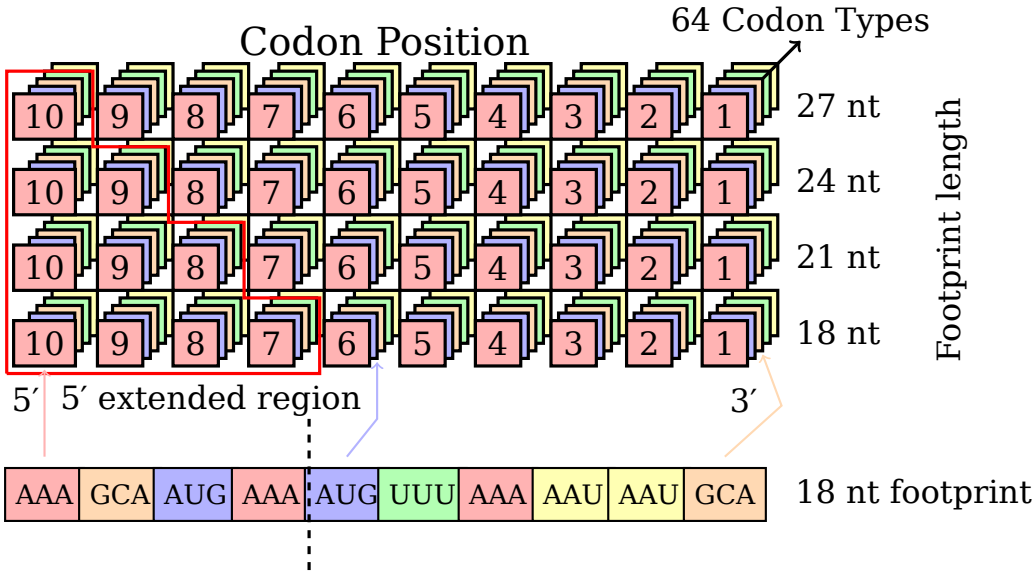


Figure 2.1: Statistical analysis of footprint codon composition. We count how many times each type of codon occurs in ribosome footprints, while keeping track of the position within the footprint as well as the length of the footprint. This information is stored in a matrix, where the color represents a type of codon and the number is the positioning away from the 3′ end of the footprint.

GSE33671) [Oh et al., 2011]. All other ribosome footprinting data for *E. coli* cells grown in 3-morpholinopropane-1-sulfonic acid (MOPS) medium come from SRR1067765, SRR1067766, SRR1067767 and SRR1067768 (GSE53767) [Li et al., 2014]. Yeast data were from cycloheximide datasets SRR1363415 and SRR1363416, untreated datasets SRR1363412 – SRR-1363414 and 3-AT datasets SRR1363420 – SRR1363423 (GSE58321) [Lareau et al., 2014] or SRR950450, SRR1040415, SRR1040423 and SRR1040427 for proline enrichment analysis (GSE50049) [Artieri and Fraser, 2014b]. Initial processing of the sequencing data was done largely as outlined previously [Ingolia et al., 2013]. In detail, sequence read archive (SRA) files were converted to FASTQ files using SRA toolkit 2.3.2-5, program fastq-dump. Ends were trimmed using FASTX toolkit version 0.0.13.2, program fastx clipper, using options `-Q33 -a NNN -l 21 -c -n -v`, where the sequence NNN was either `CTGTAGGCACCATCAAT` or `AAAAAAAAAAAAA`, depending on the library preparation method.

A Bowtie library of rRNA and tRNA sequences was compiled and Bowtie 1.0.0 was used to remove these sequences from the trimmed fastq file with options `--quiet -p 8 -l 23 --un=SRRxxx.norrna.fq rrna seqs -q SRRxxx.trimmed.fq >/dev/null` (where SRRxxx is the input file name of the SRR dataset).

The remaining sequences were aligned using Bowtie against the *E.*

coli MG1655 or *Saccharomyces cerevisiae* S288c reference genome with options `-S -p 8 -l 21 --sam-nohead genome name -q SRRxxx.norrna.fq >SRRxxx.aligned.SAM`. Any unaligned reads were discarded using GNU grep version 2.18 by keeping only those lines not matching the pattern 'XM:i:0' (`grep -v -E 'XM:i:0' SRRxxx.aligned.SAM >SRRxxx.matched.SAM`). A merged file of multiple experimental replicates was then created using GNU cat version 8.21.

The SAM file was converted to a simplified file using a program executed by Perl version 5.18.2. Briefly, sequence mismatches relative to the genomic reference were corrected, based on the reasonable assumption that the majority of mismatches are due to sequencing errors and not to biological variability. These corrected sequences were grouped into sets of identical sequences and mapping positions, keeping track of their mapped genomic positions and the number of occurrences. A Perl program was used to discard reads which were not mapped within coding regions or were within 10 codons of the ends of coding regions. For calculations involving codons, frame was maintained either by adding and/or removing 0 or 1 nt at either end, or by shifting the sequence upstream 0, 1 or 2 nt and making the length a multiple of three. Sequences were then divided into sets of 3 nt, representing the in-frame codons and the codons at each position were tallied,

keeping into account the total length of the sequence and the number of times the sequence occurred. The resulting tallies were normalized to percentages and output to a file aligned to either the 5′ or 3′ end. A similar procedure was used to compute nucleotide frequencies, except the footprints were neither divided into sets of triplets nor corrected for reading frame.

2.2.3 Computational processing of mRNA-seq sequencing libraries

SAM files were prepared from SRA files as above (SRR1067773 and SRR1067774, GEO # GSE53767) [Li et al., 2014] or (SRR950758, SRR950896, SRR951829 and SRR1040263, GEO # GSE50049) [Artieri and Fraser, 2014b]. The entire lengths of the reads were aligned to their respective genes and the mean read density across all positions was taken to represent the expression level of a gene. Background mRNA codon frequencies were computed by counting the codons per gene and multiplying by the expression level, and then summing these tallies across all protein coding genes. Unlike Artieri *et al.* [Artieri and Fraser, 2014a], codon-resolution read density corrections were not performed.

2.2.4 Hierarchical clustering of ribosomal footprint codon occupancies

Ribosome footprints were 3′-aligned and the footprint positions were summed across the length classes and normalized, such that, for any position, the sum of the values across all codons at that same position (column) equals 1. Hierarchical clustering of the data was performed using pvclust [Suzuki and Shimodaira, 2006] in R 3.1.1, command `pvclust(data, method.dist='cor', method.hclust='average', nboot=100000)`. The absolute 5′ position, and any positions 12 codons or more from the 3′ end, were ignored to avoid effects from different footprint lengths as a result of the reported 5′ end biases.

2.2.5 GC content visualization of a sequence

To visualize the GC content over a particular sequence range, first the position-specific GC content, defined as 1 for G or C and 0 for A or U, was computed for the entire sequence. Next, a Savitzky-Golay smoothing algorithm [Savitzky and Golay, 1964], with window size 27 and polynomial degree 4, was applied to the resulting data to help vi-

sually discriminate between regions of high or low GC content. The algorithm was implemented in a Perl program.

2.2.6 Visualization of footprint ends over a gene using 2D heat maps

A pre-processed SAM file, with sequencing reads corrected to the genome and identical sequences and mapping positions merged and with the number of identical reads recorded, was scanned line by line and for each read a PTT file was scanned in parallel until a gene was found whose range contained the read mapped position. If a match was found, then the two endpoints were tracked, however many times the read occurred. The end result produces a separate file, for any gene with at least one mapped read, containing the locations of the endpoints for each mapped read as well as their occurrence numbers. These pairs of endpoints were then re-arranged in a two-dimensional (2D) tab-delimited text file, the position within the file (line and tab number) representing the footprint whose 3′ and 5′ ends spanned those regions. At each position in the text file, the number of footprint reads was recorded. Places with no reads were marked with a '-,' and were considered missing when plotting with Gnuplot (version 4.6 patchlevel

3).

2.2.7 Source code

All source code is made available on the GitHub repository https://github.com/atmartens/ribosome_profiling.

2.3 Results

2.3.1 LB induces serine pauses 4 codons from the 3′ end in *E. coli*

It was previously shown that, when grown in LB medium, *E. coli* ribosome footprints are more likely to contain serine codons than expected [Li et al., 2012]. This effect was attributed to a depletion of intracellular serine, and thus of aminoacylated $tRNA^{Ser}$, due to the preferential metabolism of serine amino acids in the absence of other carbon sources. This mechanism is thought to slow ribosomes at the A site, because the low levels of aminoacylated $tRNA^{Ser}$ in the cell will cause the ribosome to pause while waiting until a cognate tRNA molecule arrives. We thus reasoned that the analysis of serine codon frequencies in ribo-

some footprints could reveal the location of the A site in *E. coli*.

To address this question, the frequencies of serine codons in footprints were calculated. As predicted, all six serine codons are enriched in ribosome footprints (Fig. 2.2, p. 65). Interestingly, we find that the enrichment is positioned relative to the 3′ end and not the 5′ end. The codon positions 3 through 5 upstream of the 3′ end are highly enriched in serine, with position 4 being the strongest. This result suggests that, in bacteria, the A site is located 4 positions from the 3′ end and that comparisons between ribosome footprints of different lengths require alignments relative to the 3′ end, but not the 5′ end.

Although the most obvious pattern in this dataset is the serine enrichment, we wondered if other codons might likewise have important position-specific biases. To address this question in an unbiased way, we used hierarchical clustering to group codons by their position-specific enrichments. As expected, we find that all six serine codons cluster together, confirming the serine effect (Fig. 2.3, p. 66) and no other codons exhibited position-specific biases as important as serine in these data.

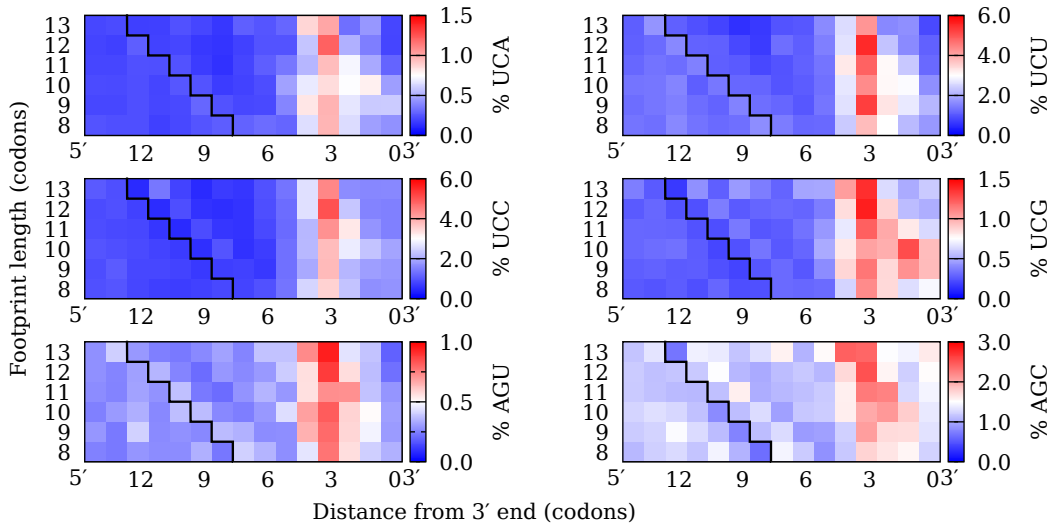
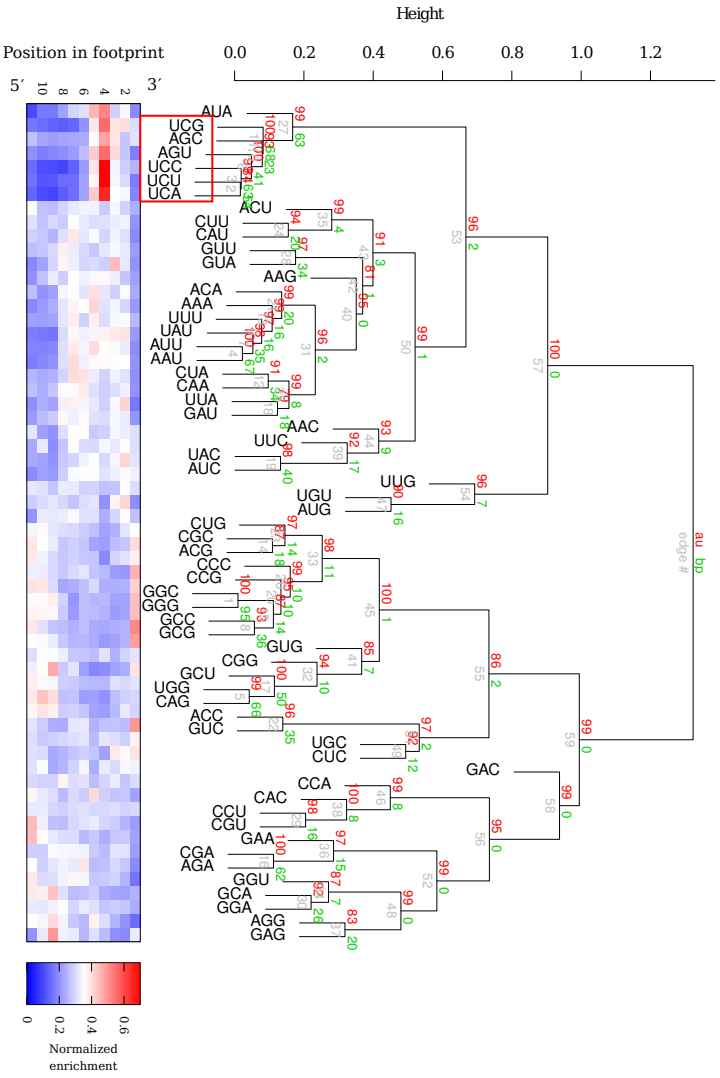


Figure 2.2: *E. coli* cells grown in LB display serine enrichment in ribosome footprints. The point of greatest enrichment is 4 codons upstream from the 3′ end, regardless of footprint length, suggesting that longer *E. coli* footprints have extra sequence at the 5′ end of the ribosome. The codons to the left of the jagged line were looked up from the genome, revealing the sequences of shorter ribosome footprints prior to nuclease digestion.

Figure 2.3: Only serine codons are enriched 4 codons from the 3′ end of *E. coli* ribosome footprints Footprints were aligned to the 3′ end, and the different length categories were merged. To determine which codons have similar enrichment patterns in ribosome footprints, hierarchical clustering was employed. After normalizing the codon levels, codons were grouped based on their position-specific enrichments. The clustering was repeated with bootstrapping to generate p-values (red numbers). We consider p-values greater than 95 to be significant.


Height
Position in footprint
0.0
0.2
0.4
0.6
0.8
1.0
1.2
5′
3′
10
8
6
4
2
AUA
UCG
AGC
AGU
UCC
UCU
UCA
ACU
CUU
CAU
GUU
GUA
AAG
ACA
AAA
UUU
UAU
AUU
AAU
CUA
CAA
UUA
GAU
AAC
UUC
UAC
AUC
UUG
UGU
AUG
CUG
CGC
ACG
CCC
CCG
GGC
GGG
GCC
GCG
GUG
CGG
GCU
UGG
CAG
ACC
GUC
UGC
CUC
GAC
CCA
CAC
CCU
CGU
GAA
CGA
AGA
GGU
GCA
GGA
AGG
GAG
au bp
edge #
0
0.2
0.4
0.6
Normalized
enrichment

2.3.2 Prolines can induce pausing in both yeast and bacteria

The incorporation of proline, an imino acid, into a nascent chain is potentially conformationally troublesome due to steric and other effects [Pavlov et al., 2009], a property which should affect all domains of life, as evidenced by the existence of special elongation factors for proline in both eukaryotes and bacteria [Ude et al., 2012, Doerfel et al., 2012, Gutierrez et al., 2013]. This snaring effect has both been measured in vitro, using stop-flow kinetics [Pavlov et al., 2009] and using ribosome profiling in yeast [Artieri and Fraser, 2014a]. Given that this effect is robust, we asked whether it could be used to calibrate the pause location in footprints of different lengths, thus giving an indication of where the ribosome P site was relative to the footprints.

We tallied the occurrences of proline codons at all positions in ribosome footprints, separated by footprint length (Fig. 2.1, p. 57) and normalized these frequencies by the background frequencies from a parallel mRNA-seq experiment [Artieri and Fraser, 2014a]. The resulting data can be visualized as a 2D heat map, where hotter colors represent enrichment in ribosome footprints relative to the mRNA background. We confirm that, as previously described in yeast, all four pro-

line codons (CCN) are enriched (Fig. 2.4, p. 70). Overall, prolines are about 1.5–2.5-fold enriched, with CCA being the most and CCG the least (Fig. 2.4). Furthermore, we find that longer ribosome footprints show similar codon frequencies as shorter footprints, if the footprints are aligned to the 5′ end. The positions which are most enriched in prolines are conserved between the lengths in vertical patterns. For example, proline codon CCA is most highly enriched 4 codons downstream of the 5′ end and enrichment is less at the other positions. Since proline incorporation is thought to retard ribosome processivity at the P site, during peptidyl transfer [Pavlov et al., 2009], we infer that the greatest enrichment at position four indicates the P site location.

Curiously, a similar analysis of *E. coli* data does not reveal this trend. We find that, overall, proline frequencies in ribosome footprints, at any position and of any length, are similar to those in the mRNA background (Fig. 2.5, p. 71), suggesting either that proline incorporation in *E. coli* is not slow or that averaging all footprints together might mask the effect. We posited that individual *E. coli* genes, containing long proline repeats, do experience ribosome pauses. For example, the *E. coli* gene *amiB* has the longest stretch of consecutive proline codons (eight). Contrary to the absence of observed pausing for individual prolines, we do find that repeat prolines coincide with a dramatic ribosome pause (Fig. 2.6, p.

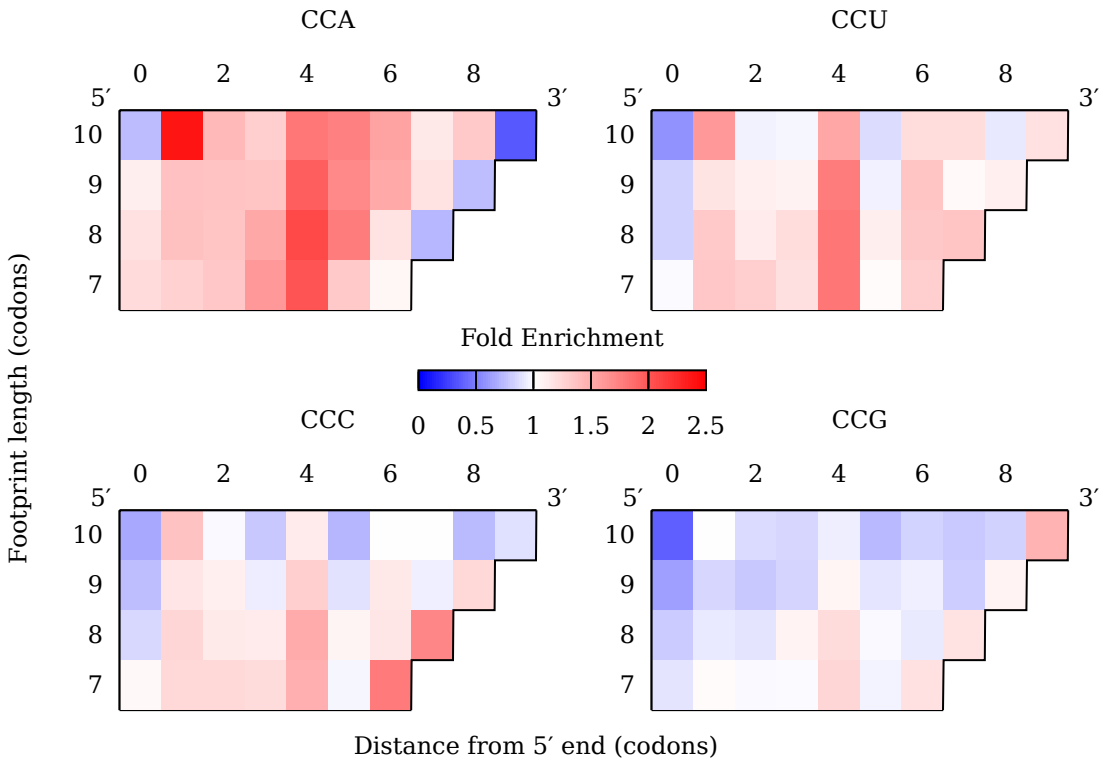


Figure 2.4: Prolines are enriched in yeast ribosome footprints. Comparing the frequencies of proline codons in mRNA to those in footprints, we find that all four proline codons are over-represented. In particular, the fourth position downstream from the footprint 5′ end is the most enriched, regardless of the length of the footprint, suggesting that yeast footprints align naturally to the 5′ end. Since ribosome stalling during proline incorporation takes place during peptidyl transfer, these data suggest the P site is located 4 codons downstream of the 5′ end in yeast.

72), suggesting that, at least in extreme situations, prolines can also induce ribosome stalling in *E. coli*.

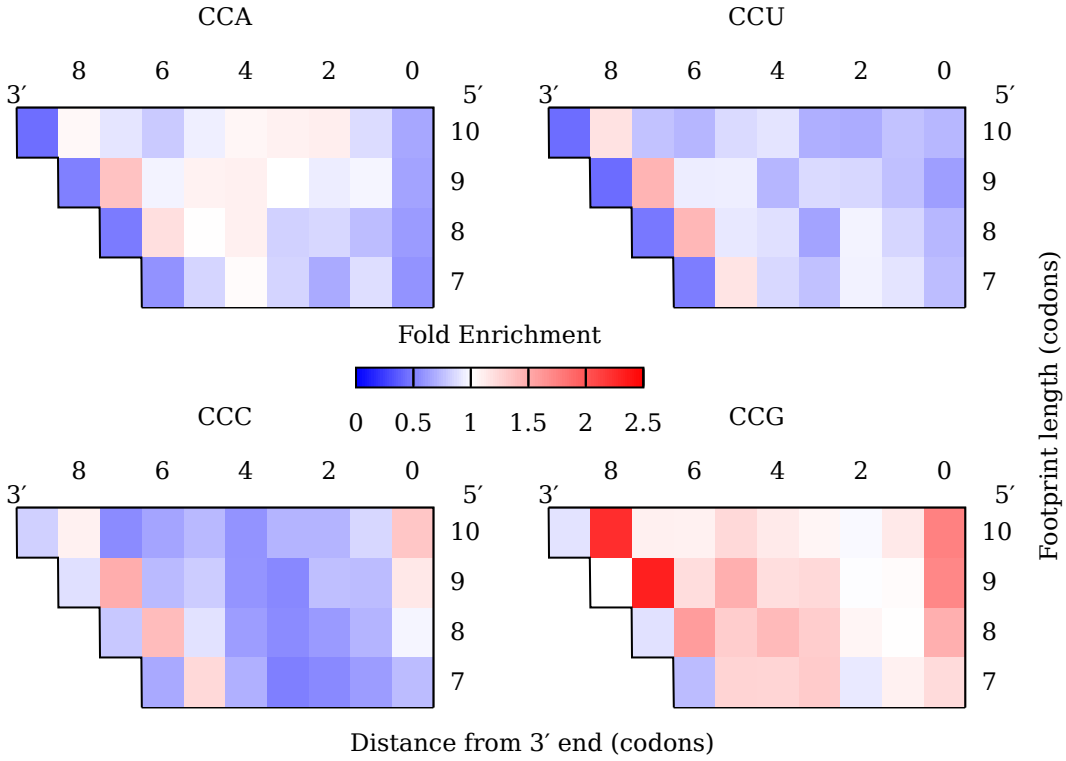


Figure 2.5: Differential proline enrichment in *E. coli*. Although the proline codon CCG, and to a lesser extent CCA, is enriched in E. coli footprints, the codons CCU and CCC are not.

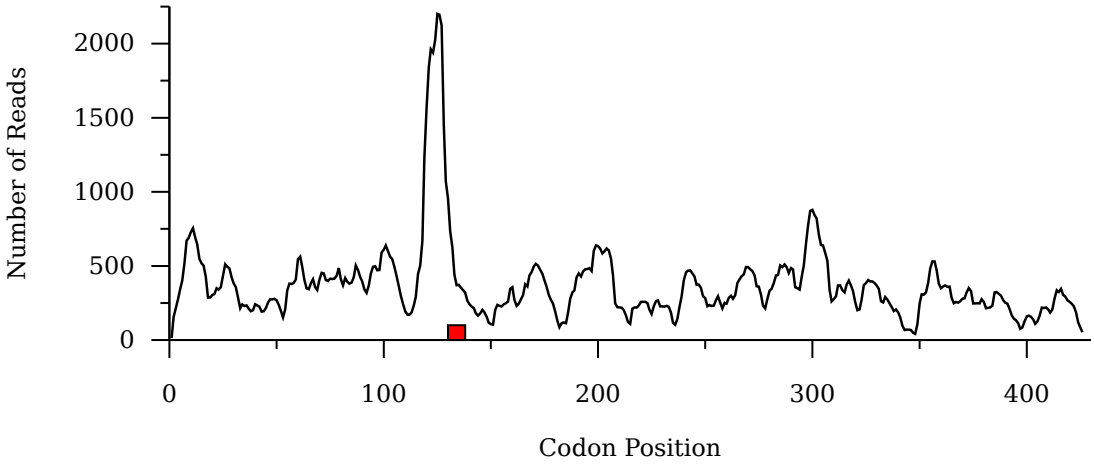


Figure 2.6: Ribosome density is higher in a proline-rich sequence. The E. coli gene *amiB* has eight consecutive proline codons (box, codons 130 to 138) which coincide with a sharp rise in ribosome footprint density, suggesting that, at least in extreme cases, proline residues do retard translation.

2.3.3 3-AT causes histidine enrichment at all positions

The small molecule 3-AT (3-amino-1,2,4-triazole) was previously shown to specifically inhibit the incorporation of histidines into nascent proteins [Guydosh and Green, 2014, Lareau et al., 2014], presumably by reducing the intracellular $tRNA^{His}$ concentrations [Klopotowski and Wiater, 1965]. The reduction in tRNA exclusively results in in the enrichment of histidine codons, but not other codons, in ribosome footprints [Lareau et al., 2014]. Since a reduction in tRNA will increase the time a ribosome waits at the A-site, whereto tRNA molecules diffuse randomly, we reasoned that a comparison of the histidine codon frequencies in ribosome footprints would reveal the location of the ribosome A site.

As before, the codons were tallied at all positions and categorized by ribosome footprints of different lengths. As shown previously, we find that histidine codons are enriched after 3-AT treatment, at nearly all positions, comprising a much higher fraction of the total number of codons in ribosome footprints than from untreated cells (Fig. 2.7, p. 75). Although the effect is not limited to a single position, the enrichment at the position 5 codons downstream from the 5′ end is slightly stronger than at other positions. Despite the weak signal, this position is the

likeliest candidate of the A site, for it agrees with the proline data, which set the P site 4 codons downstream the 5′ end. Aligning the footprints to the 5′ end, as before, tends to conserve vertical enrichment patterns across footprints of different lengths, again suggesting that, in yeast, the additional mRNA in longer footprints extends from the 3′ end of the ribosome.

2.3.4 GC-rich 5′ sequences are present both in yeast and *E. coli*

It was previously shown that *E. coli* and *Bacillus subtilis*, but not yeast, ribosome footprints have GC-rich 5′ ends [Li et al., 2012] and that this effect correlates with footprint length [O'Connor et al., 2013]. This effect has been explained by the interactions between G-rich mRNA and the C-rich rRNA (3′ -CACCUCCU-5′), called the aSD effect, and is a ribosome pausing mechanism which happens away from the A, P or E sites. Given the previously discussed pausing effects, which do take place at the A or P sites, we asked how these different mechanisms contribute to ribosome pauses as measured by ribosome profiling experiments.

To address this question, we tallied the frequencies of all codons, both by position and by footprint length. The aSD model predicts that

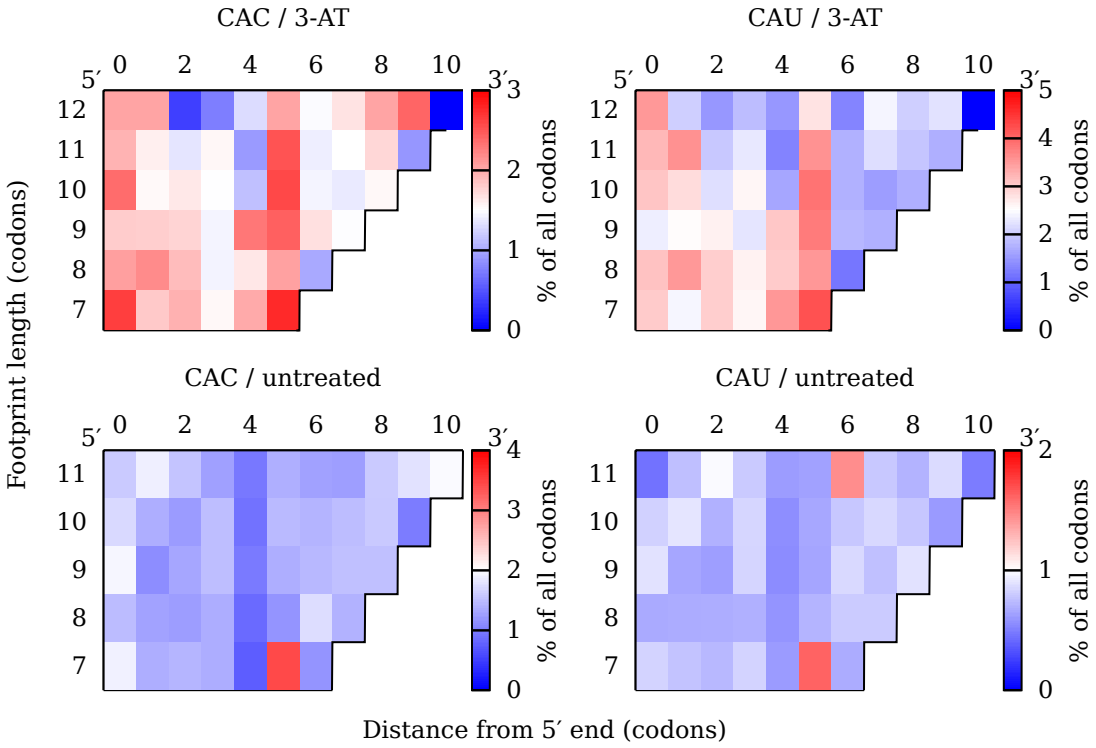


Figure 2.7: Histidine codons are enriched following 3-AT treatment. The depletion of intracellular $tRNA^{His}$ using 3-AT causes a sharp increase in histidines in ribosome footprints (top) compared to untreated cells (bottom). Unlike the proline enrichment, which had a clear position-specific effect, this enrichment is more evenly spread across the length of the footprints. However, a strongly conserved vertical pattern of enrichment at position 5 might indicate the A site position.

G-rich codons will be more present than other codons, specifically at the 5′ ends, in a length-dependent manner and will not be enriched in yeast. To see what sequences are 'missing' from the shorter footprints, these were extended to 45 nt by looking up the sequences from the genome and aligned to the 3′ end. According to the 'inch-worm' hypothesis [O'Connor et al., 2013], the aSD model predicts that longer footprints are able to fit inside the ribosome if they bind the rRNA more strongly, giving rise to a correlation between footprint length and 5′ G content. This hypothesis also predicts that, in *E. coli*, the sequence occurring immediately prior to 'shorter' footprints will not be high G, and that in yeast there should be no difference between the nucleotide content of the longer or shorter footprints.

We confirm that *E. coli* ribosome footprints 5′ ends are G- rich and that this enrichment grows with footprint length (Fig. 2.8, p. 78). Furthermore, the extended 5′ regions of the shorter footprints are not enriched, supporting the aSD model. However, we notice that the extreme 5′ end position is GC-poor, an observation not predicted by the aSD model. In addition, an analysis of yeast ribosome profiling data also reveals GC-enriched sequence at the 5′ ends of longer footprints (Fig. 2.9, p. 79), which was not expected. We wondered if some of these nucleotide biases could be due to artifacts from the library preparation

or sequencing processes. Nucleotide analyses of mRNA-seq data show how the library-generation process for both mRNA-seq and ribosome profiling do cause nucleotide-level GC biases, but that these 5′ end GC enriched sequences are distinct from the biases observed in the mRNA-seq data (Fig. 2.10, p. 80). Therefore, a process specific to ribosome profiling likely imparts specific nucleotide biases in the sequencing data.

2.3.5 Visualizing footprint densities as a function of read length

The relationships between sequence composition and footprint length demonstrate the importance of accurately visualizing footprint density across transcripts. Given the potential for ambiguity in the comparison of footprints of different lengths, we have developed a graphical method for simultaneously displaying footprint density, along a transcript, of all different lengths. Importantly, this method allows us to interrogate the relationships between ribosome density and mRNA sequence of individual genes, thereby avoiding the problems associated with genome-wide averaging of the sequencing data, as shown above.

For example, we highlight a region within the *E. coli* gene *rpsA*, which encodes a highly-expressed protein component of the small ribosome

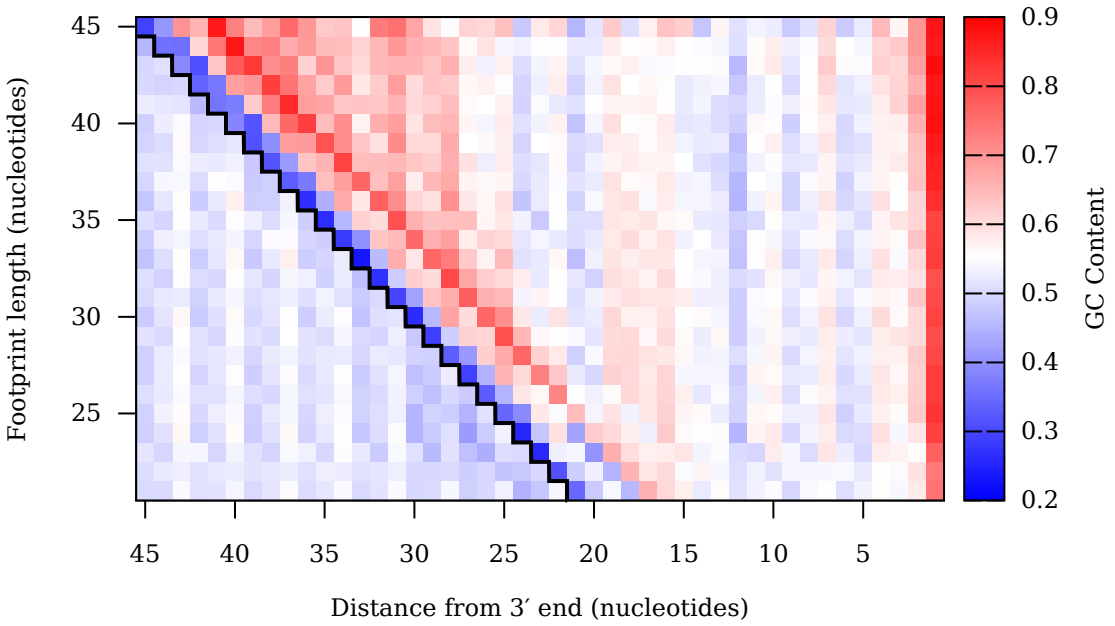


Figure 2.8: GC content analysis of *E. coli* ribosome footprints. Footprints were categorized by length and each nucleotide position was measured for its average GC content across all footprints. Shorter footprints were extended to the 5′ end, revealing what the mRNA sequence was prior to nuclease digestion (jagged black line). We see that the 5′ and 3′ ends of footprints, regardless of length, have some strong GC biases. Furthermore, the longer footprints have increasingly high GC content downstream from the 5′ end.

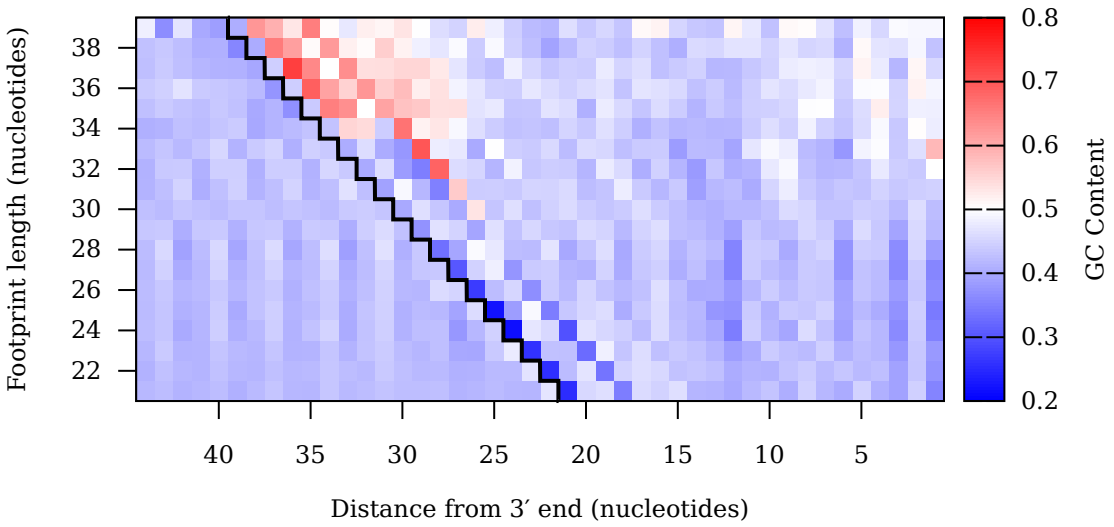


Figure 2.9: GC content analysis of yeast ribosome footprints. The yeast genome has a lower GC content than the *E. coli*, as visible by the cooler colors. Nevertheless, we observe, as in *E. coli*, sequence biases at the ends of the footprints as well as a length-dependent increase in GC content downstream from the 5′ end.

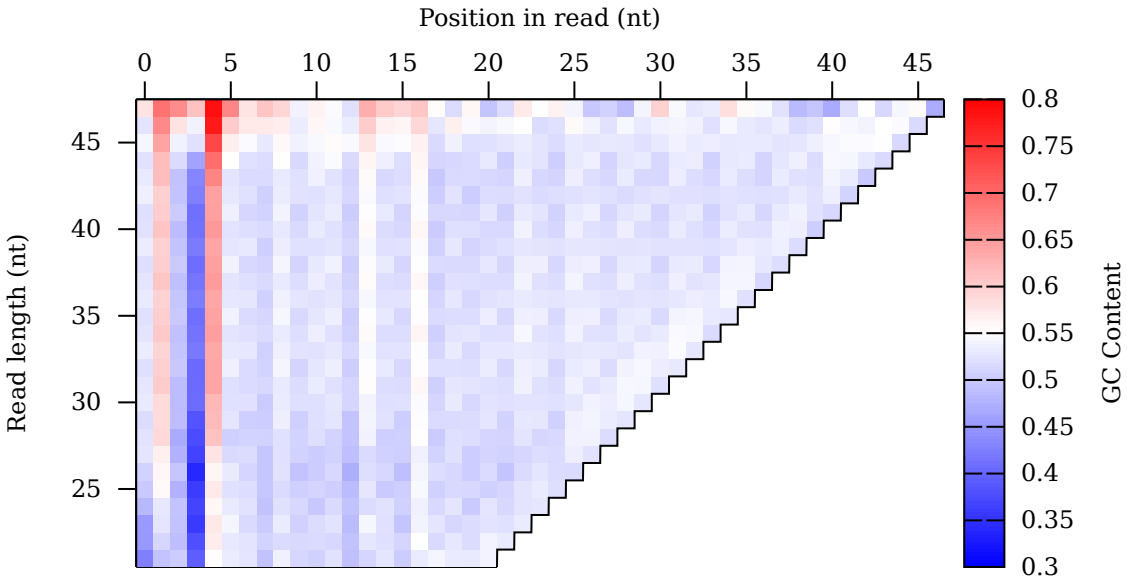


Figure 2.10: GC content analysis of mRNA seq data. To better understand the sequence biases noticed in the ribosome footprinting data, they were compared to mRNA-seq data, which underwent many of the same steps but fundamentally report on different biological phenomena. We find that these data have nucleotide biases near the 5′ ends of the reads, but that there is little to no GC enrichment in a length-dependent manner.

subunit. By plotting the footprint coordinate endpoints on the x and y axes, and the read number as a heat map, we can see how footprint length heterogeneity varies across a gene (Fig. 2.11, p. 82). Remarkably, we find that regions of low footprint density correlate with low GC regions, exactly as seen in the previous analysis of total ribosome footprints. The nucleotide biases at the ends are also apparent: footprints tend to be GC-poor at either extremity, regardless of length. Similarly, using this method we can see how the peak in read density along the *E. coli* gene *amiB* (Fig. 2.12, p. 83) coincides with the eight consecutive proline codons and how these footprints have a slightly altered length profile compared to those from the rest of the gene, a feature which would have remained hidden with more conventional plots.

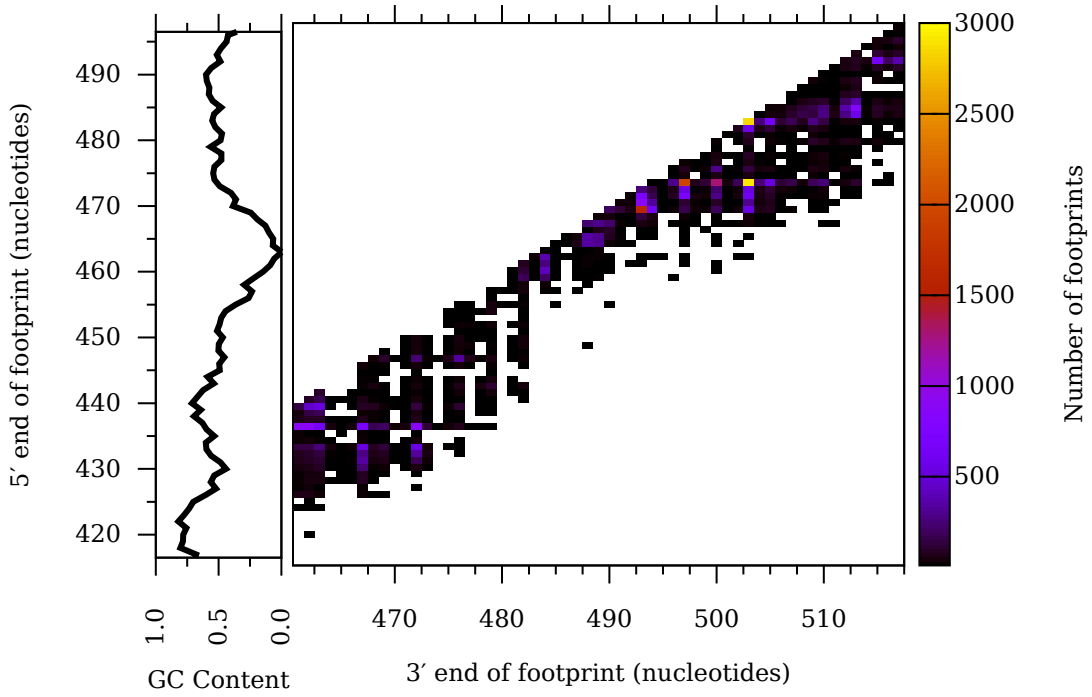


Figure 2.11: Plotting ribosome footprints by tracking the 3′ and 5′ end coordinates and displaying the read count using a heat map. A section of the highly-expressed *E. coli* gene *rpsA* shows areas of high and low sequence coverage, indicating which parts of the transcript have, on average, more or less ribosome density. Using the footprint length information juxtaposed with the average GC content, we can see how there is a strong relationship between the read length and the 5′ GC content.

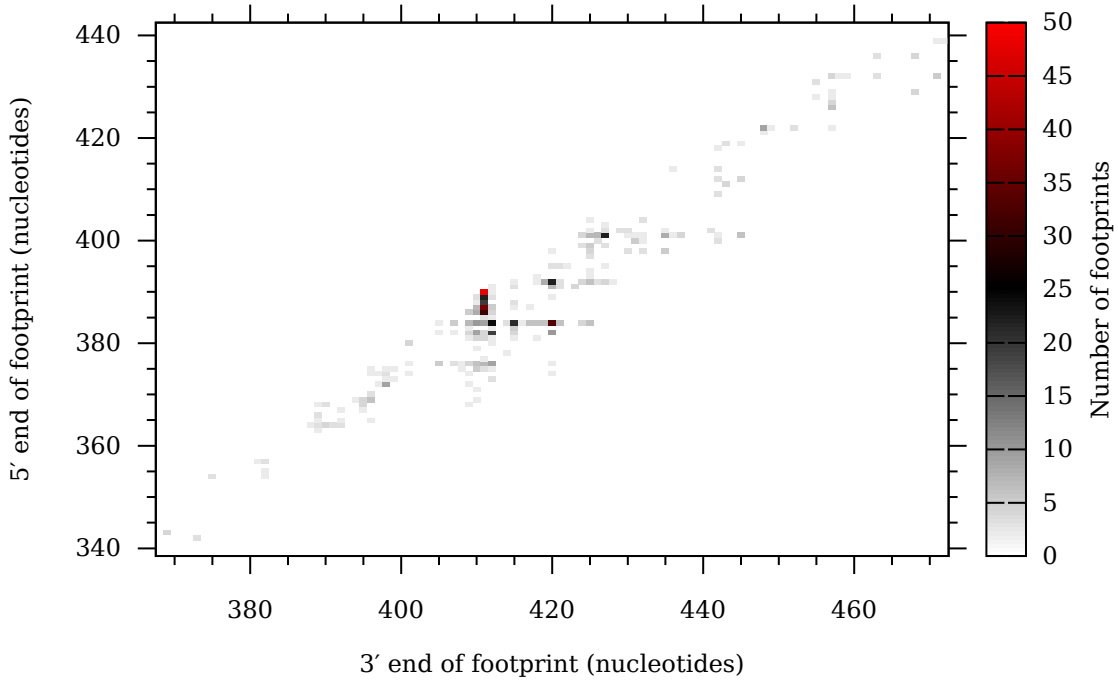


Figure 2.12: Plotting ribosome footprints by tracking the 3′ and 5′ end coordinates and displaying the read count using a heat map. The lower expression level of the gene *amiB* is reflected by a lot of empty space, but clearly the peak of reads at the eight consecutive prolines stands out.

2.4 Discussion

2.4.1 Understanding footprint length differences through position-specific codon frequencies

The principle underlying the ribosome profiling technique is the protection of mRNA from nuclease by the ribosome. Though a typical footprint is 28 nt long, recent evidence has demonstrated that many footprints are longer or shorter. For example, ribosomes which halt at truncated mRNA molecules lack sequence at their 5′ ends, producing ~15 nt footprints [Guydosh and Green, 2014]. Footprint length heterogeneity has also been associated with different conformations of the ribosome, which can be preferentially selected through the use of small molecule inhibitors of translation elongation [Lareau et al., 2014]. Furthermore, the distribution of sequencing read lengths has been used to distinguish between true ribosome footprints and other nucleic acids, such as mRNA or rRNA which is protected from nuclease treatment by non-ribosomal molecular complexes and which therefore do not constitute true ribosome footprints [Ingolia et al., 2014]. The mechanism by which elongation inhibitors generate footprint length heterogeneity is

unknown and these length differences complicate the analysis of ribosome footprinting data: *a priori*, there is no default way to compare two footprints of different lengths.

Without any further information, it is conceivable that footprints should be aligned to the 5′ end, to the 3′ end, somewhere in between, or that no single alignment is correct. We have empirically found that aligning yeast footprints to the 5′ end preserves the same position-specific codon frequency differences between the different lengths and agrees with previously reported alignment methods in yeast [Lareau et al., 2014]. Crucially, two steps of translation, which take place at either the A or P sites and are located one codon apart, are found to cause codon enrichments which are also one codon apart. The distinct mechanistic explanations for these two types of pauses agree with our knowledge of ribosome structure. This agreement is particularly compelling given that the experiments were peformed independently and confirms that ribosome profiling is sensitive enough to detect these types of translation rate differences.

This alignment places the ribosomal P site 4 codons downstream of the 5′ end of the footprint and the A site 5 codons away. Longer footprints thus appear to, on average, have extra mRNA at the 3′ end. *E. coli* footprints, however, align to the 3′ end, as demonstrated by the sharp

serine enrichments 4 codons upstream from the 3′ end; the explanation of this difference remains unclear, but suggests that, unlike in yeast, additional mRNA extends past the 5′ end of the mRNA and that different computational methods need to be applied when examining eukaryotic or bacterial datasets.

2.4.2 Sequencing longer footprints to distinguish artifacts from biological mechanism

In addition to the aforementioned position-specific codon biases in footprints, which we can attribute to specific steps during protein synthesis, we also find, as previously reported [Artieri and Fraser, 2014a, Hansen et al., 2010], that sequencing data contain substantial GC biases, only some of which are shared between mRNA-seq data and ribosome profiling data. Although these biases are problematic because they complicate the estimates of true molecular levels, in principle it should be possible to partially correct for over- or under-representation of ribosome footprints by using the mRNA-seq data as a reference [Artieri and Fraser, 2014a]. However, we also notice biases present in the ribosome profiling data, both in yeast and in *E. coli*, which are absent from

the mRNA-seq data. These biases could either be due to library generation steps unique to ribosome profiling or to true translation events. For instance, the 5′ ends of footprints are GC-rich, while the ends of mRNA sequencing reads show no such bias. This type of 5′ enrichment was previously attributed to the aSD effect in *E. coli* [Li et al., 2012, O'Connor et al., 2013], but the same effect is also present in yeast, calling into question the hypothesis.

The efficiencies of several *in vitro* processing steps, such as linker ligation, reverse transcription or polymerase chain reaction, could be impacted by terminal sequence composition. However, these three steps were also used when generating the mRNA-seq data. The defining step during ribosome profiling experiments, which is absent from other sequencing experiments, is the treatment of the sample with nuclease [Ingolia et al., 2009]. Given that nucleases tend to cut GC-poor sequences [Hörz and Altenburger, 1981, Dingwall et al., 1981], rather than GC-rich sequence, we propose that longer ribosome footprints result from incomplete nuclease digestion at mRNA 5′ ends. This was hinted at using experiments looking at ribosome stacking on a single bovine transcript, where lesser nuclease treatment revealed longer footprints [Wolin and Walter, 1988], and could easily be tested by performing ribosome profiling experiments with variable nuclease treatments. If nuclease sen-

sitivity explains footprint length heterogeneity, it would also suggest that different ribosome conformations, such as those induced by small molecule inhibitors of translation [Lareau et al., 2014], have different nuclease susceptibilities. Likewise, the length differences of 'contaminant' sequences could be explained by variable nuclease susceptibility in these other molecular complexes [Ingolia et al., 2014].

The possible role library generation plays in introducing sequence biases highlights some key steps of the ribosome profiling methodology which must be carried out carefully, lest the data become skewed or misinterpreted. Although a canonical footprint is 28 nt in length, in practice we observe footprints as long as 45 nt. This means that sequencing reads should be at least 50 nt long and that the gel purification of size-selected fragments should likewise include longer molecules. Omitting these precautions will result in the under-representation of GC-rich sequences in the final datasets and might explain why GC-rich 5′ footprint ends were not noticed in the original yeast studies.

This chapter was previously published in Nucleic Acids Research under the title "Ribosome A and P sites revealed by length analysis of ribosome profiling data.", [Martens et al., 2015], https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv200, and contains some minor changes.

Chapter 3

An integrated study of translation rates using ribosome profiling and stochastic simulations

3.1 Introduction

Pronounced translation kinetics differences, in particular at ribosome pause sites along an mRNA, have been implicated in downstream biological processes such as the regulation of gene expression and of co-translational protein folding. Yet measurements of the translation elon-

gation rate of specific mRNAs are surprisingly few. The *in vivo* translation rate of the *E. coli lacZ* mRNA, a model for gene regulation, has been determined using several different methods, consistently ranging between 10 to 15 a.a s^{-1} under typical conditions [Kepes and Beguin, 1966, Lacroute and Stent, 1968, Schleif et al., 1973, Sørensen et al., 1989]. The most precise measurements of the *lacZ* translation rate were determined using radioactively labeled methionine pulse-chase experiments [Sørensen et al., 1989]. It was estimated that the translation rate averages 12.5 a.a. s^{-1}, with some evidence of minor variations across the mRNA. Given this information, it should be expected that ribosome pausing along the *lacZ* mRNA is relatively uniform.

Although the relationship between the translation elongation rate and the accumulation of ribosome density is straightforward, it has been shown using simulations that the consequences of ribosome queue formation are not trivially understood (reviewed by [von der Haar, 2012]). Simulating translation has been done using "top-down" approaches, which use differential equations to derive mathematical relationships between physical parameters, and "bottom-up" approaches which begin with rules for ribosome movement and follow individual trajectories over time. In order for these simulations to successfully predict ribosome densities, accurate parameters must be provided. This information, including the

translation initiation and elongation rates, is unknown for most genes, however the *lacZ* mRNA has been extensively characterized (*e.g.* [Kennell and Riezman, 1977, Sørensen et al., 1989]).

It is now possible to use ribosome profiling to investigate the ribosome density across an mRNA. Yet, using this technique to measure genome-wide translation rates has not yet been achieved. For example, ribosome profiling can detect pauses at specific types of codons which are slower than average [Subramaniam et al., 2014, Martens et al., 2015, Woolstenhulme et al., 2015], yet the degree to which ribosome profiling reports on translation rates is not well understood. We previously collected ribosome profiling data of *E. coli* which, incidentally, were expressing *lacZ*, numbering 30 experiments in total (Ch. 4, p. 136). By comparing the resulting ribosome density measurements with previous measurements of the translation rate and stochastic simulations we provide new evidence that the translation rate of the *lacZ* mRNA is not constant.

3.2 Methods

3.2.1 Simulation of a polysome

The simulation begins with a naked mRNA. After each tick of the global clock a ribosome is either loaded onto the mRNA, or progresses by 1 codon, if a new pseudorandom number (between 0 and 1) is less than or equal to the probability for the start codon or the A-site codon. New ribosomes are placed onto codons 0 – 4, with the A-site codon located at position 1 [Martens et al., 2015, Woolstenhulme et al., 2015]. Each ribosome stores its A-site positioning as well as its 5′ and 3′ coordinates, which are increased by one upon successful advancement. The ribosome's footprint size initially grows as the ribosomes progresses until it reaches a maximum length of 10 codons. For each iteration, first the most advanced ribosome (nearest the 3′ end) is processed, followed in order by those successively nearer the 5′ end. However, a ribosome is never allowed to load onto the mRNA or progress forwards by one codon if its 3′ coordinate is blocked by a downstream, neighboring ribosome's 5′ coordinate. When a ribosome terminates translation it is entirely removed in a single step. The entire trajectory history for all ribosomes can be recorded (Fig. 3.1, p. 94).

The simulation requires as input the length of the mRNA, the transi-

tion probabilities for each codon and the total simulation time. It then proceeds in discrete time steps until reaching the given maximum number of iterations. This algorithm is similar to others in the literature [Mitarai et al., 2008]. The master clock ticks once per iteration and is unitless — its physical meaning depends on how the transition probabilities are defined. We use the convention that a probability of 0.1 is equivalent to a rate of 20 a.a. s^{-1}, which means that the fastest possible rate is 200 a.a. s^{-1} ($p = 1.0$). This maximum rate is roughly ten times greater than typical rates in *E. coli* and therefore satisfies the required time resolution. If desired, other conversion factors can be used instead, balancing resolution with execution time.

Multiple mRNAs can be simulated in the same experiment. Either a fixed number are generated at the same initial time, or new ones are generated with some frequency. To speed up this procedure, each polysome is spawned as a new thread. These threads are "synchronized" by initiating them with the iteration number at that spawn time. All code was executed using Perl v5.24.1.

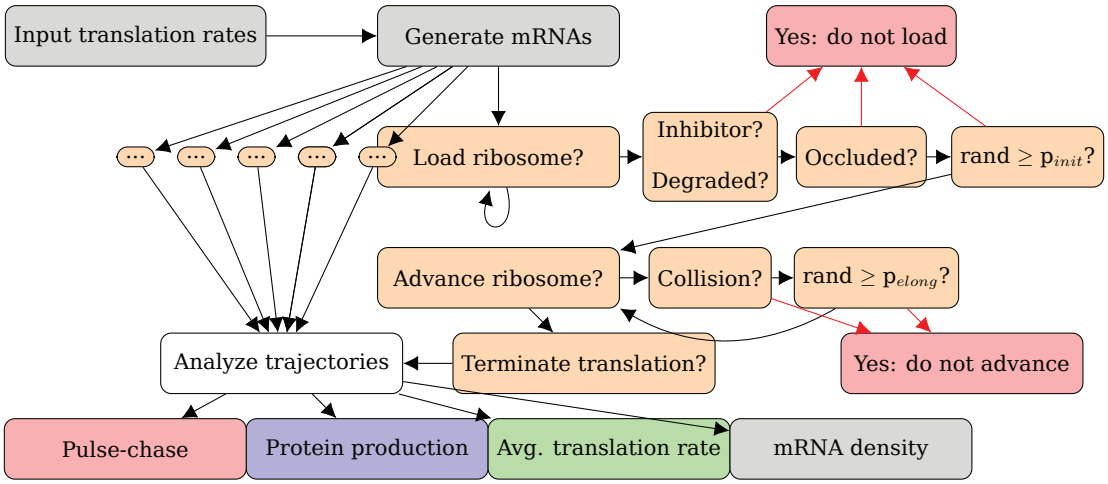


Figure 3.1: The logical flow of the polysome simulator. Using a set of translation rates, one for each codon, many polysomes can be simulated, providing detailed access to the microscopic state of the system at each discrete time interval. The individual ribosome trajectories can then be analyzed to re-construct the macroscopic state, including simulating the radioactive pulse-chase experiment, observing the increase in protein over time, calculating the average translation rate and calculating the mRNA density.

3.2.2 Simulating production and pulse-chase labeling of new proteins

The polysome simulation allows for not only the generation of individual ribosome trajectories over time, but also for the calculation of global properties of all polysomes in the simulation. In particular, the rate of protein production following inducer, as well as the incorporation of radioactively labeled amino acid, can be calculated from these simulations. To measure the protein production over time as following the addition of inducer, each iteration a new mRNA is generated with $p = 0.005$ (0.1 s^{-1}). Each polysome records the number of completed proteins at each iteration value, and these values are then summed across all polysomes to compute the total increase in proteins over time. In the case of an *in vitro* translation system to which a large quantity of mRNA are added at once, the simulation is performed with a fixed number of mRNA at the outset. As with the delay time assay, the data are the summed proteins per iteration across all polysomes.

The pulse-chase experiment also begins with a fixed number of mRNA, which are allowed to populate with ribosomes and reach a steady state. In practice, this means that this experiment can be performed at the same time as the "*in vitro* delay time" experiment. After the simulation

is complete, the individual trajectories of the ribosomes on each mRNA are analyzed. This step also requires the input of the pulse start time and duration, the amino acid sequence of the protein and the type of amino acid to label with. Any trajectory which eventually completes a protein during the experiment and was being synthesized during the pulse of labeled amino acid is included in the analysis. This strategy allows for probing the same experiment using different combinations pulse and chase times or of labeled amino acids without re-simulating the polysomes.

For each trajectory the two ribosome A-site positions, at the start (pulse) and end (chase), are determined. If the protein came into existence after the pulse but before the chase, or was released from its ribosome before the chase and after the pulse, then these values are used instead. Then, the equivalent amino acid residues corresponding to these two codon positions are calculated, and the number of labeled residues within this window is determined from the amino acid sequence. This value is added to an array containing a number for each iteration, representing which nascent peptides were completed and when. In the following examples only methionine was labeled, though any other combination of labeled amino acids can be used as well.

3.2.3 Experimental measurement of the translation rate using the delay time assay

The delay time assay was performed as in [Schleif et al., 1973], with the following modifications. A 250 ml flask containing 15 ml MOPS glycerol medium was inoculated with *E. coli* MG1655. After overnight growth in the exponential phase (250 RPM, 37°C, doubling time of 90 min) and reaching an OD_{600} of ∼0.40, the cells were transferred to a pre-warmed 100 ml beaker, shaking at 37°C, and incubated for 1 min. IPTG (GoldBio, cat. no. I2481C) was added (t_{zero}) to a final concentration of 1 mM, and 120 µl aliquots were removed at 5 s time intervals into 1.5 ml Eppendorf microcentrifuge tubes containing 180 µl MOPS medium with 550 µg/ml chloramphenicol. To consistently remove the aliquots at exactly 5 s, a timer was placed next to the test tube rack and set to count down. Cells were slowly removed by micropipette for two seconds, to prevent the formation of air bubbles, and then were injected into the tubes in less than one second. Once all aliquots were removed they were incubated 10 min at 37°C in order to allow newly-synthesized proteins to fold. 300 µl of buffer 1 (200 mM $NaPO_4$ pH 7.0, 2 mM $MgSO_4$, 200 mM β-mercaptoethanol) was added, followed by 5

µl of toluene and vortexing for 5 s. 120 µl of buffer 2 (4 mg/ml ONPG (*ortho*-Nitrophenyl-β-galactoside, Sigma Aldrich cat. no. 73660) in 50% buffer 1) was then added successively to each sample, and the reactions were incubated 1 h in a 28°C water bath. The reactions were then quenched with 300 µl of 1 M Na_2CO_3. 240 µl were transferred to a clear 96-well plate (Falcon, cat. no. 353072), and A_{420} measurements were performed using a Berthold TriStar2 plate-reader luminometer with a 420 nm emission filter (Berthold, cat. no. 39452).

The absorbance at 420 nm, which reports on the concentration of ortho-nitrophenol, is proportional to the level of β-galactosidase enzyme. The total enzyme level depends on the basal expression before induction (b) and the rate of increase in mRNA and enzyme levels following activation of transcription and translation (m). However, enzyme levels only increase after the delay (τ), which represents the time for the following steps to occur: 1. entry of inducer into the cell, 2. de-repression of LacI, 3. transcription of the ribosome binding site, 4. translation initiation and 5. synthesis and release of a complete monomer by the ribosome. An incubation step allows for the formation of functional β-galactosidase tetramers, which is not considered part of the delay. To model the transition before and after the appearance of new enzyme, we used the heavyside unit step function H. Before this delay, only basal

enzyme is present, and m and τ are ignored ($t \leq \tau$, $H(t-\tau) = 0$). After the delay, m and τ are included ($t > \tau$, $H(t-\tau) = 1$) (Eq. 3.1, p. 99).

$$A_{420}(t) = b + m \cdot (t-\tau)^2 \cdot H(t-\tau) \tag{3.1}$$

3.2.4 The theoretical protein production rate as a function of the initiation and elongation rates

The mean time $\bar{\tau}$ required for a ribosome to engage and then leave the start codon is the sum of the mean times of these two steps (Eq. 3.2, p. 99). The time required for initiation is just the inverse of the initiation rate k_i. However, the time for clearing space for the next ribosome depends on the amount of mRNA sequence occupied by a ribosome, w. If $w = 10$ codons, then the time required for leaving this space is the inverse of the mean elongation rate of the first 10 codons, k_e, multiplied by 10.

$$\bar{\tau} = \frac{1}{k_i} + \frac{w}{k_e} \tag{3.2}$$

3.2.5 Cloning the Luciferase gene into the high copy pUC19 vector

In addition to the vectors described in chapter 4.2.2, p. 145, an additional vector bearing the gene encoding fast luciferase was assembled for reasons beyond the scope of this chapter. Cells expressing this vector were also induced with IPTG, and their β-galactosidase ribosome profiling results have been included in this analysis. The P_{lac} promoter, Luc F CDS and downstream transcriptional terminators were amplified by PCR using primers ATM46 (5′-CCCGAAAAGTGCCACCTGACTGGCACGACAGGTTTCCC-3′) and ATM47 (5′-GCCTTTTGCTGGCCTTTTGCTCAAAGAGTTTGTAGAAACGCAAAAAGG-3′), with Q5 DNA polymerase, 10 mM primers, 0.2 ng / µl template (pACYC177 / lucF) and the following parameters: 98°C denaturation for 30 s, and then 35 cycles of 98°C for 10 s, 64.5°C annealing for 15 s, 72°C extension for 75 s, and a final 2 min denaturation at 72°C. pUC19 DNA was digested with restriction enzymes ZraI (New England BioLabs, cat. no. R0659) and PciI (New England BioLabs, cat. no. R0655). The PCR product and vector backbone were ligated using the Gibson assembly method and transformed into *E. coli* NEB 5α cells (cat. no. C2987). After plating onto LB Ampicillin plates and overnight growth at 37°C, colonies were picked and re-streaked. Plasmid DNA was prepared us-

ing the GeneJet miniprep kit (ThermoFisher cat. no. K0502), and was then transformed into *E. coli* MG1655.

3.2.6 Ribosome profiling

Ribosome profiling was performed as in chapter 4.2.5, p. 152. The footprints across the *lacZ* mRNA were assigned using constant 30 nt widths, aligned to the 3′ end of the read. Then, total densities were window averaged with a window size of 201 nt.

3.2.7 Translation Index

The same translation index was used as described in chapter 4.2.1, p. 139.

3.3 Results

3.3.1 *lacZ* ribosome densities are not uniform

To collect ribosome density data for the *lacZ* mRNA, we performed 30 different ribosome profiling experiments in *E. coli* grown under the

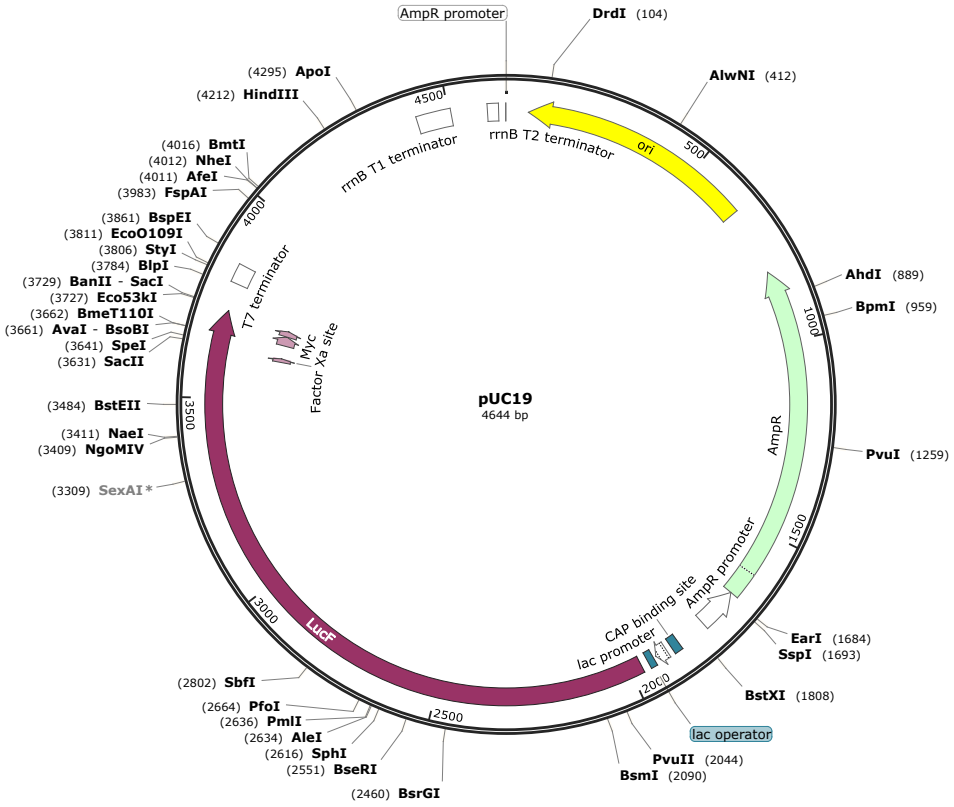


Figure 3.2: The pUC19 vector bearing Luc F. The Luc F gene was cloned into this high-copy vector using Gibson assembly on the strand opposite the ampicillin resistance gene *bla*. Cells carrying this pasmid were induced with IPTG, activating *lacZ* exprssion, and analyzed using ribosome profiling.

same conditions, in minimal medium with IPTG. We note that these *E. coli* strains are not identical, as they each bear one of several different pACYC177 plasmids, some of which carry synonymously re-coded Luc genes (for more details, see Ch. 4, p. 136). Since we have not found any evidence for changes, between these strains, in position-specific ribosome pausing along endogenous mRNAs (data not shown), we inferred that Luc expression does not affect the translation kinetics of the *lacZ* mRNA. We do observe that ribosome densities are not uniform along the mRNA in all the individual profiles, however inferring statistical significance from sequencing experiments depends on both the number of replicates and their sequencing depths [Busby et al., 2013], and so it is not possible to draw conclusions from any single dataset. Therefore, we took advantage of the high sequencing depth and statistical power provided by such a large number of experimental samples to investigate ribosome pausing as measured using ribosome profiling.

After merging these datasets, we observe that ribosome densities, as measured using ribosome profiling, are not uniform (Ch. 3.3, p. 105). For comparison the translation index [Spencer et al., 2012], a dimensionless scale which has been used as a proxy for translation rates (Ch. 4.2.1, p. 139), also does not yield a uniform profile. Since neither profile is uniform, we wondered whether the ribosome profiling data indicate

local ribosome pauses which had not previously been detected. Such pauses could conceivably be the result of codon-dependent translation rate differences, or due to collisions between or queuing among neighboring ribosomes.

To explore these possibilities we have developed a software program which stochastically simulates the progression of ribosomes along an individual mRNA. By analyzing many individual ribosome trajectories together, it is then possible to simulate the types of experimental observables which have previously been reported in the literature. In addition, information which is not easily measured from experiments, such as the frequency, duration and location of collisions between adjacent ribosomes, can also be gathered. In these simulations, discrete rates for ribosome movemement are programmed for initiating at the start codon, as well as moving from one codon to the next, with codon-specific resolution. However, these intrinsic rates are theoretical upper bounds which are only attained at low ribosome densities, as ribosome binding and movement may also be limited by collisions with other ribosomes.

3.3.2 Simulating the “delay time” assay

There are several experimental techniques which measure the translation rate of an mRNA. The delay time between addition of inducer

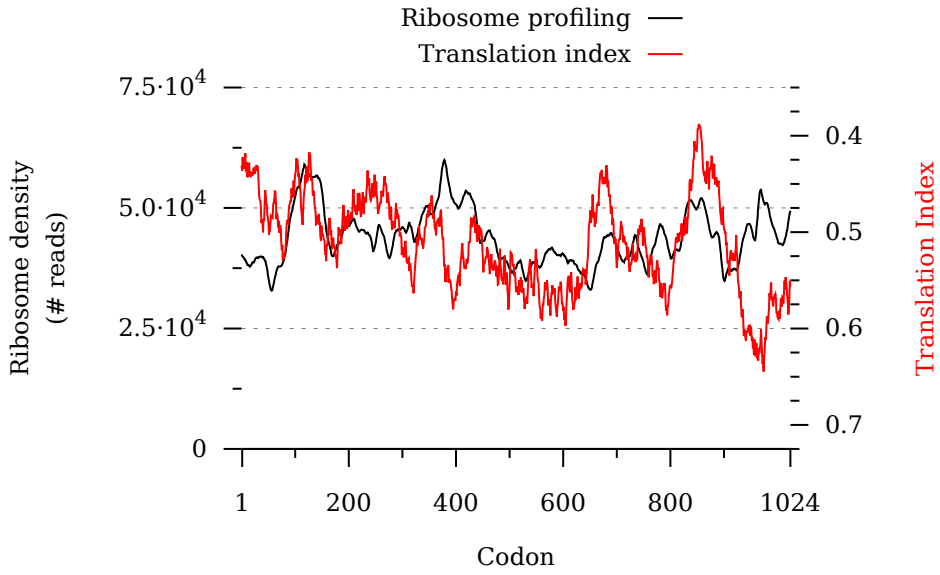


Figure 3.3: Ribosome densities and predicted translation efficiencies along the *lacZ* mRNA are variable. Ribosome occupancies in *E. coli* cells expressing *lacZ* were recorded, producing a high level of sequencing depth and 30 separate replicates. For comparison, the average predicted translation efficiencies are also shown [Spencer et al., 2012] (Ch. 4.2.1, p. 139; Table 4.1, p. 145). Both the sequencing data and the translation efficiency exhibit considerable variation across the mRNA. This variation could indicate local differences in average translation rates.

(IPTG) and the appearance of new β-galactosidase protein is often used to determine the average translation rate of an mRNA (Ch. 1.5, p. 24). We can simulate this experiment by generating many independent mRNAs and tabulating the time to completion of the individual proteins. These mRNAs are randomly generated over time, as would occur following induction of transcription by the inducer. The linear increase in mRNA copy number over time, combined with translation of these mRNAs, produces a quadratic increase in protein over time [Schleif et al., 1973]. We find that the simulated time of appearance of new protein clearly depends on the translation rate (Fig. 3.4, p. 107), and produces a curve with the same characteristics as produced experimentally (Fig. 3.5, p. 108), confirming that the simulation is able to recapitulate known behavior.

3.3.3 The *lacZ* translation rate is not uniform

It was previously observed that the *lacZ* mRNA translation rate is not constant [Sørensen et al., 1989]. This conclusion was reached using the radioactive pulse-chase labeling of methionine residues within nascent β-galactosidase. This experiment works by adding short pulses

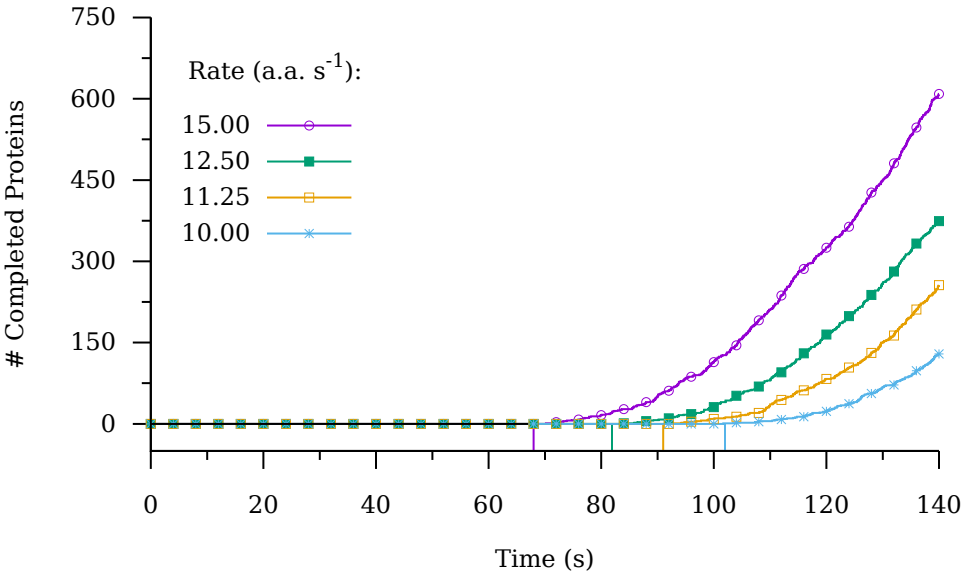


Figure 3.4: The delay time between addition of inducer and appearance of the first new proteins is affected by the translation elongation rate. The polysome simulator was used to compare how the production protein rate depends on the translation elongation rate. For each simulation, new mRNAs were randomly generated over time, producing the expected linear increase in the levels of transcript and a quadratic increase in protein levels. The translation rate controls the time before the appearance of the first proteins, but has little effect on the subsequent rate of protein production.

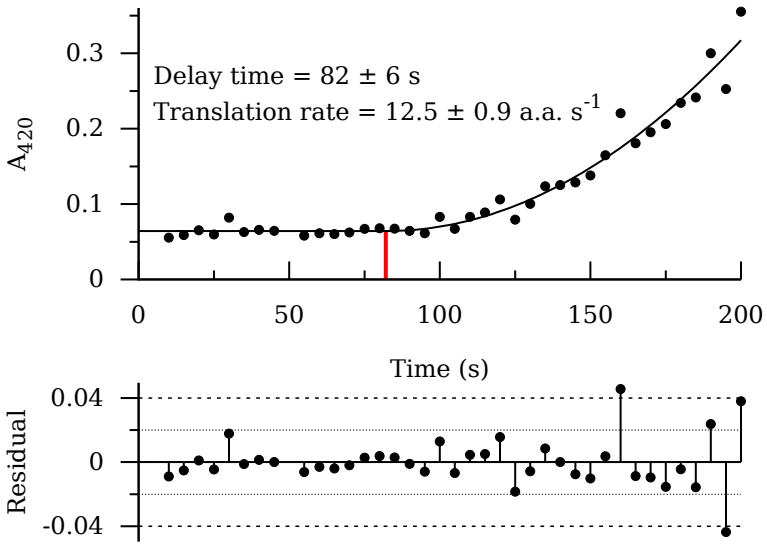


Figure 3.5: Our experimental translation rate measurement of the *E. coli lacZ* mRNA agrees with the literature values. The first noticeable change in β-galactosidase activity occurs at roughly 100 s after the addition after IPTG, however fitting the model from Eq. 3.1 indicates that the first new β-galactosidase proteins were synthesized after 82 s. This time corresponds to a 12.5 a.a. s^{-1} translation rate, in good agreement with other measurements [Kepes and Beguin, 1966, Lacroute and Stent, 1968, Schleif et al., 1973, Sørensen et al., 1989].

of radioactively labeled methionine to a growing culture with active ribosomes. Then, after a variable amount of time, the radioactive methionine is diluted (chased) with non-radioactive methionine. Proteins which were completed between the pulse and the chase contain radioactivity if a ribosome incorporated a labeled methionine residue during that time. These proteins accumulate increasing amounts of radioactivity until the time difference between the pulse and chase exceeds the amount of time required to incorporate all methionine residues (which is approximately the same as the time required to make a full protein), at which point the radioactivity is maximal (and equivalent to every methionine being labeled). To estimate the translation rate of the *lacZ* mRNA, we simulated the incorporation of methionine residues in nascent proteins using two different, constant translation rates along the mRNA. As expected, the shape of the resulting incorporation profile is similar to what has been experimentally determined (Fig. 3.6, p. 111). Yet, neither simulation fits the entire curve. In particular, the last ~20 sec of timepoints stand out as being different from the first ~70 sec, consistent with a shift in the translation rate. By carefully comparing different regions of the various simulated curves to the experimental data, we iteratively refined a combination of two translation rates which reproduces the entire curve. This result predicts that translation begins at a rate of 9.0 a.a. s^{-1} before

proceeding to 13.5 a.a. s^{-1}. However, testing this prediction is difficult, as further refining the translation rate with smaller time intervals is technically challenging and not trivial.

Using these translation rates, we next tallied the final ribosome densities on 10,000 *lacZ* mRNAs simulated for 280 s (Fig. 3.7, p. 112). As expected, the resulting density profile has higher density downstream of the start codon, where translation is slower (codons 1-251), followed by lower density, where translation is faster. However, this profile does not strongly resemble the ribosome profiling data that we have collected, suggesting that there is more local variability in the *lacZ* mRNA translation rate than is readily measured using the pulse-chase method.

3.3.4 Stochastic ribosome progression promotes ribosome collisions

Ribosome pausing along the *lacZ* mRNA, caused by translation of slow codons, could be accentuated by collisions or queuing among adjacent ribosomes. Queue formation, or the close association between consecutive ribosomes, slows down translation whenever a ribosome blocks the progression of other ribosomes. The degree of queue formation depends on the relative magnitudes of the translation initiation

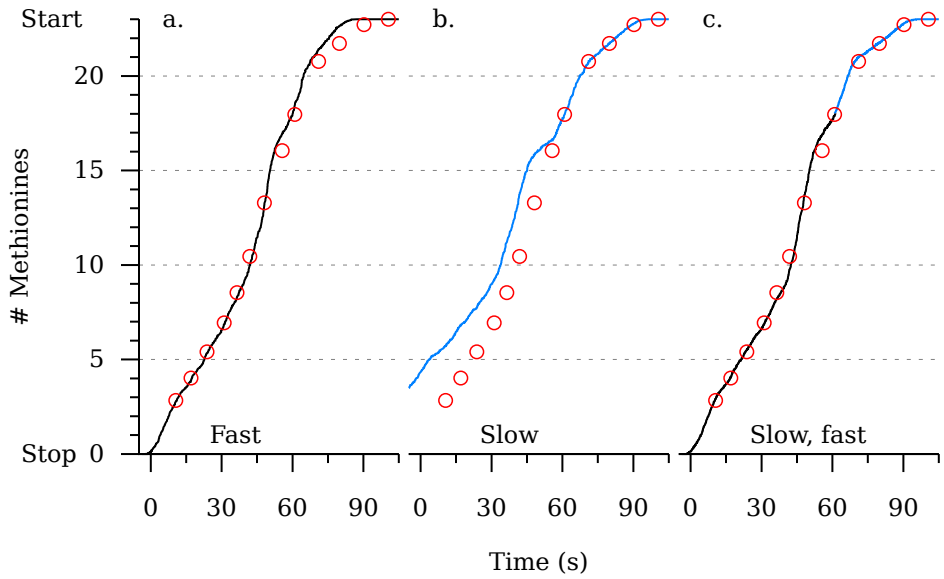


Figure 3.6: *lacZ* mRNA is not translated with a uniform translation rate. After allowing the simulated ribosomes to populate mRNA for 100 s, the nascent proteins were labeled with radioactive methionine, followed by a chase of unlabeled methionine. a. Constant fast (13.5 a.a. s^{-1}) translation. b. Constant slow (9.0 a.a. s^{-1}) translation, with an offset to show the overlap at later timepoints. c. Slow translation for the first 250 codons (blue) and then fast for the remainder (black). This merged simulation gives a result similar to literature values across the entire simulation (red circles, from [Sørensen et al., 1989]). Slight deviations can be seen around 45 and 60 s, which could either indicate noise in the original measurements or slight deviations from the fast rate. These rates yield an overall average of 12.1 a.a. s^{-1} which is only slightly lower than previously estimated.

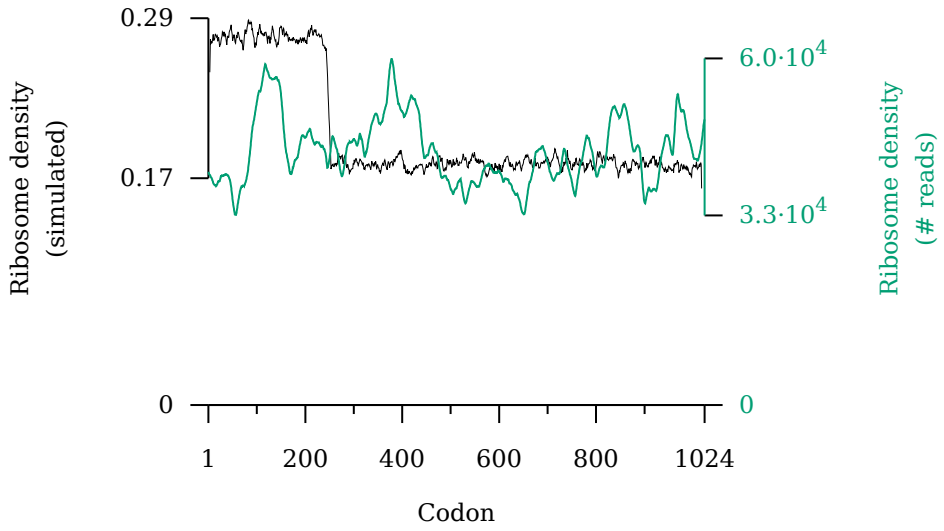


Figure 3.7: The *lacZ* ribosome density is more complex than predicted from pulse-chase experiments. Ribosome density along the *lacZ* mRNA as simulated by translation at 9.0 and 13.5 a.a. s^{-1}. 10,000 mRNAs were simulated for 280 s, and the final densities of each simulation were averaged together.

and elongation rates. Queueing is analogous to cars traveling on a single lane road, entering a toll bridge [Itano, 1968]. Much as the flow of traffic becomes congested when many cars slow down to pass through the toll gates, ribosomes form queues when the rate of adding new ribosomes to a polysome is much greater than the rate of ribosome progression along the mRNA. To evaluate the likelihood of queue formation on the *lacZ* mRNA, we simulated different combinations of translation initiation and elongation rates and looked for several signatures.

We began by simulating translation using the estimated parameters in the literature. Translation of the *lacZ* mRNA was simulated for 90 s (Fig. 3.8 p. 114), slightly more than the time for the completion of a β-galactosidase protein when translating at 12.5 a.a. s^{-1}. The initiation rate was set to 0.33 s^{-1}, as has been previously reported [Kennell and Riezman, 1977, Sørensen et al., 1989]. We can clearly see the occasional, random appearance of ribosomes at the 3′ end of the mRNA. These ribosomes progress at roughly uniform rates, as evidenced by the approximately parallel trajectories. However, even though many short-lived collisions are observed, suggesting that ribosome collisions are possible given this combination of initiation and elongation rates, no significant queues appear to form.

If ribosome progression is stochastic, then the synthesis times of a

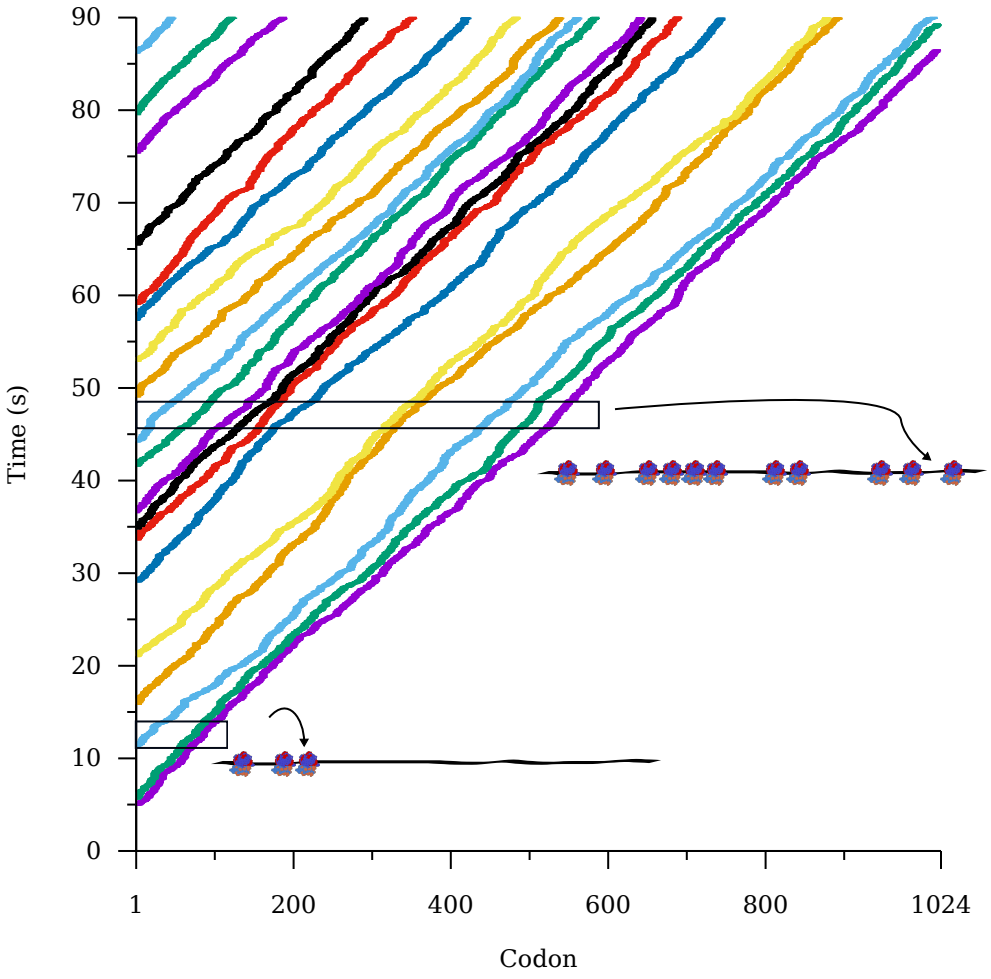


Figure 3.8: Stochastic simulations provide detailed insight into the behavior of ribosomes on an mRNA. The average translation rate is constant across the mRNA, yet there are both collisions and large gaps between adjacent ribosomes. Initiation and elongation rates were 0.33 s^{-1} and 12.5 a.a. s^{-1}, as previously measured for the *lacZ* mRNA [Kennell and Riezman, 1977, Sørensen et al., 1989, Sørensen and Pedersen, 1991].

protein will vary from one ribosome to the next. While strictly constant translation initiation and elongation rates produce a defined spacing between consecutive ribosomes, stochastic ribosome progression can amplify the formation of ribosome queues by increasing the probability of ribosome collisions. Attempts were previously made to show that the translation rate is heterogeneous using experiments performed on *E. coli* [Talkad et al., 1976]. Inducer (IPTG), transcription inhibitor (rifampicin) and translation initiation inhibitor (kasugamycin) were alternately added to actively translating *E. coli* cells. Following the addition of IPTG, RNA polymerase initiates the synthesis of new *lacZ* mRNA, but the total level of activation is kept minimal by the subsequent addition of rifampicin. β-galactosidase activity increases as a result of the translation of these new mRNAs. Addition of rifampicin then causes the β-galactosidase activity to reach a plateau, as the initiation of ribosomes is also prevented. Importantly, this transition was gradual, not abrupt, suggesting that some ribosomes took longer than others to complete translation. However, these experiments are confounded by the possibility that other steps prior to the translation of the mRNA are also heterogeneous, including the diffusion of the IPTG, rifampicin, and kasugamycin as well as the transcription by RNA polymerase. By simulating the translation of mRNA before and after the addition of a trans-

lation initiation inhibitor, we can count the accumulation of completed proteins over time. And, unlike in experiments, we know for sure that the only source of heterogeneity in the simulation is the probability of ribosomes initiating or advancing. Indeed, we observe effects similar to those previously reported (Fig. 3.9, p. 117). Both the appearance time of the first completed proteins, and the plateau of the last completed proteins, are dispersed, and this dispersion is directly attributable to the heterogeneous translation rates (Fig. 3.10, p. 118).

3.3.5 The translation initiation and elongation rates determine protein production rates

For values for the rate of translation initiation and for elongation step times in physiologically expected ranges, the rate of protein production can sometimes depend on both how often ribosomes initiate translation and the time required to complete a protein. In general, the translation initiation rate is thought to dominate [Andersson and Kurland, 1990], however relatively slow elongation rates which promote queue formation might also have an effect. To investigate the effects of ribosome collisions on protein production we varied the translation initiation or

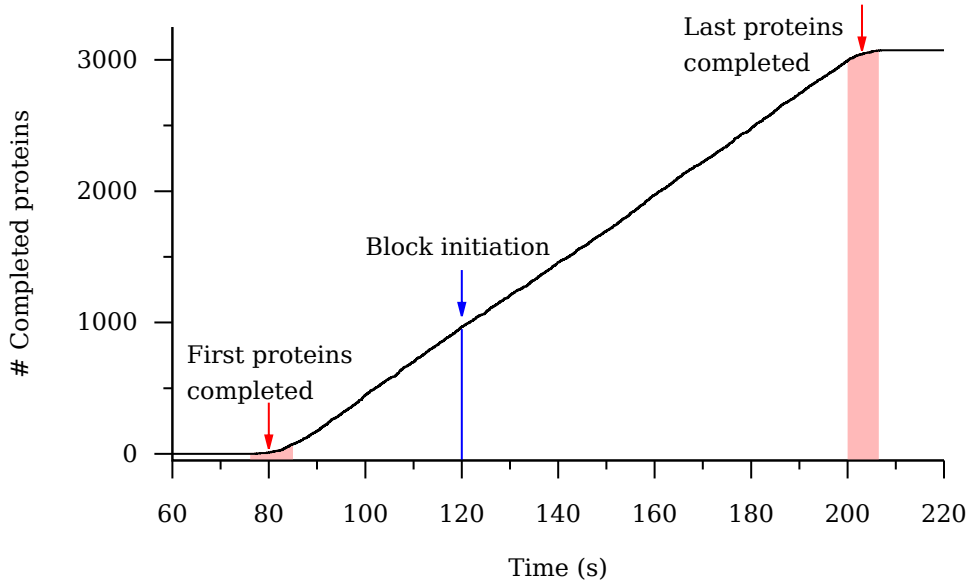


Figure 3.9: Stochastic translation leads to detectable differences in ribosome speeds. Simulations were performed by generating a fixed number of mRNAs and then inhibiting translation initiation at 120 s (blue line). If every ribosome were to have the same translation rate, then the two transitions, at 81 and 202 s, would be identical across all mRNAs (shaded areas). However, these transitions are more gradual, as expected if some ribosomes are faster or slower than average. Similar results have previously been observed [Talkad et al., 1976].

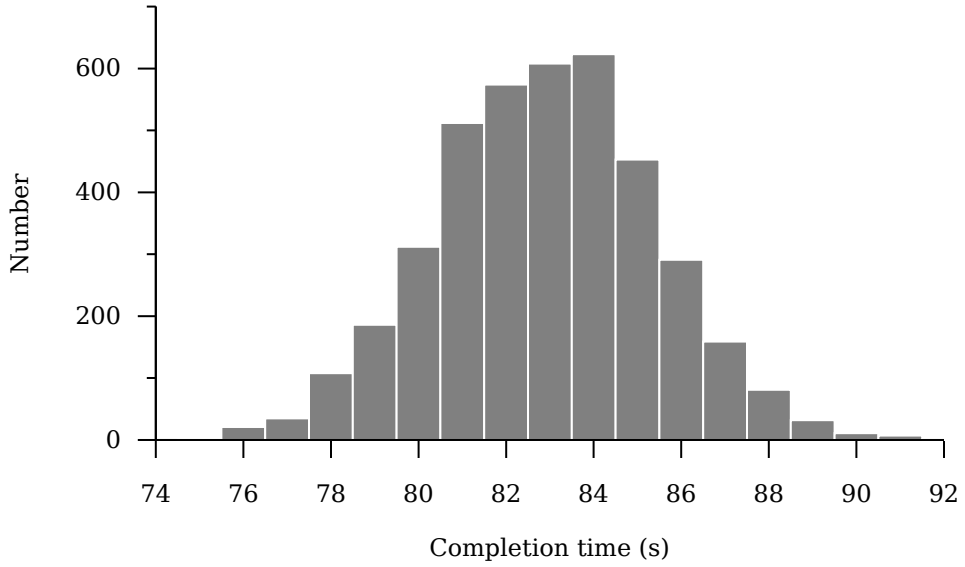


Figure 3.10: Simulated protein completion times are heterogeneous. The time elapsed between the appearance and subsequent removal of a ribosome from the mRNA was recorded for all ribosomes. Although the intrinsic translation parameters are equal for all ribosomes, and all codons, variability in the pseudorandom number generator leads to a distribution of completion times consistent with truly random and independent probabilities.

elongation rates (the termination rate was not considered). We then simulated translation of a fixed number of mRNAs with the same starting time and recorded the production of protein over time. This is similar to an in vitro translation experiment to which a large quantity of mRNA is added at once (*e.g.* [Yu et al., 2015]).

The resulting functions have two characteristic parameters: the slope, which depends primarily on the translation initiation rate, and the intercept, which depends on the elongation rate. As expected, for the same sets of parameters used previously we find that the appearance times are the same, with faster translation leading to shorter times (Fig. 3.11, p. 120). Since the initiation rates are the same, the slopes of the lines are also the same, indicating that the production rates are unaffected by the elongation rate. Interestingly, these lines actually are slightly curved immediately following the appearance of the first proteins. In our simulations this curvature is due to ribosomes which translated faster than average (as discussed previously). Experimental data with similar features have been previously published, indicating that this effect might be real [Dalbow and Bremer, 1975].

The translation elongation rate clearly affects the appearance time of new protein. Conversely, it might be expected that the initiation rate will not affect the appearance time, but only the net synthesis rate of the

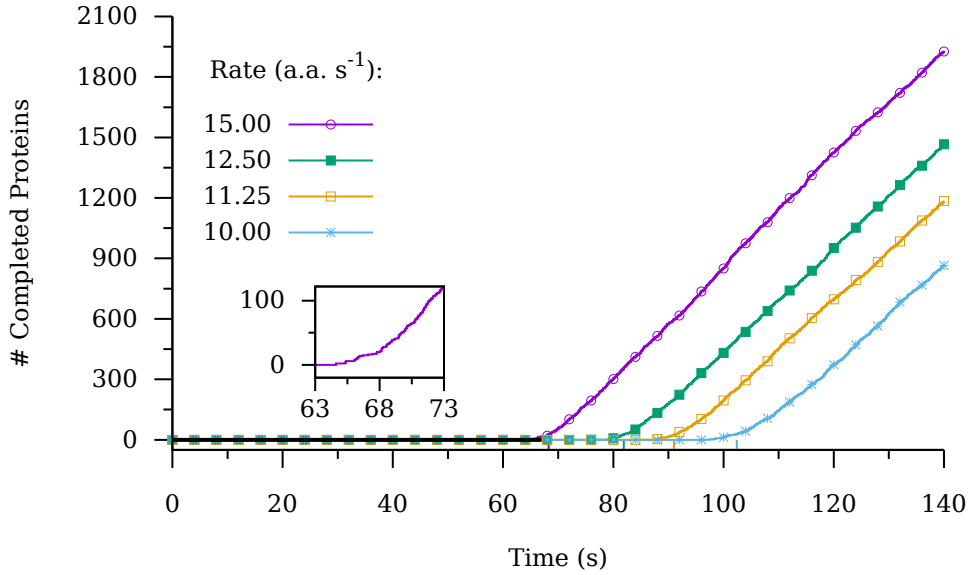


Figure 3.11: Protein synthesis rates are not affected by the translation elongation rate when initiation predominates. The kinetics of protein production were simulated as before, except that a constant number of mRNAs was generated at time zero. After the appearance of the first complete proteins, the protein production rate is nearly linear and not strongly affected by the elongation rate. The appearance function is curved before becoming linear. This initial curvature represents the appearance of ribosomes which were translating faster than the average. Four simulations are presented, along with the expected appearance times (vertical bars). Similar data have been previously described [Dalbow and Young, 1975].

protein [Andersson and Kurland, 1990]. To evaluate the importance of translation initiation rates we compared appearance simulations with different initiation rates but the same elongation rate. We clearly observe that the resulting slopes are affected by the initiation frequency, but that the appearance times are nearly indistinguishable (Fig. 3.12, p. 122). When combining this behavior with the effect of the translation elongation rate, we find that simultaneously changing the translation initiation and elongation rates can lead to complex behavior. Changing the elongation rate can create an apparent effect on the protein production rate, as the amount of protein produced over time passes through a "crossover" point. (Fig. 3.13, p. 123). This sort of behavior has possibly been misinterpreted in the literature [Yu et al., 2015].

3.3.6 A negative feedback loop between translation initiation and elongation limits ribosome collisions

Increasing the translation initiation rate leads to a corresponding increase in the protein production rate, yet the production rate reaches a maximum value. To better understand this effect we compared the observed initiation rates, calculated as the actual probability of an initi-

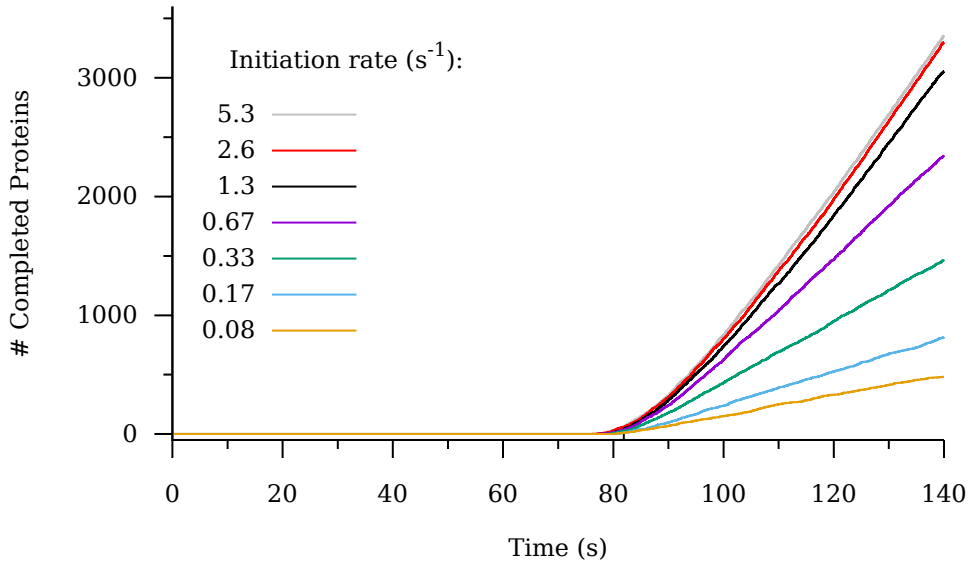


Figure 3.12: Translation initiation rates affect protein production rates. Several different initiation rates were sampled for mRNA translated at a constant 12.5 a.a. s^{-1}. When initiation is relatively infrequent then this rate affects the appearance of protein over time (yellow, blue, green, purple lines). However, when the initiation rate is limited by the elongation rate then increasing the initiation frequency has a negligible effect (black, red and grey lines).

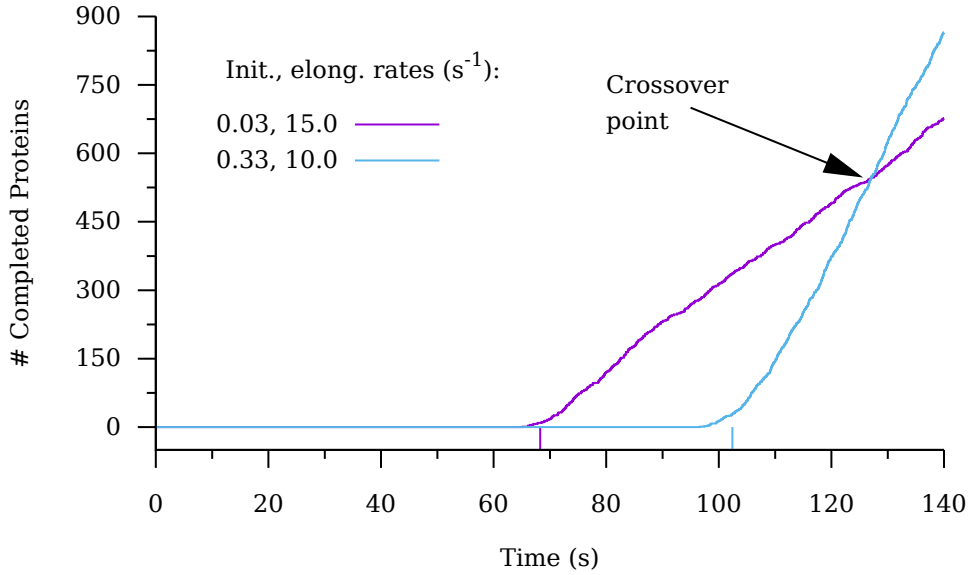


Figure 3.13: Protein production rates depend on both translation initiation and elongation rates. The protein production rates in two separate simulations, with different translation initiation and elongation rates, are compared. At first, fast translation elongation has a greater effect than fast initiation on protein production rates. However, the net amount of protein produced by slower, but more frequent, ribosomes eventually outpaces the translation by fast, infrequent ribosomes. Changing these two parameters can lead to a "crossover" point, the time at which the total amount of protein synthesized is equal. These two parameters could be modulated to produce different regimes of protein synthesis rates.

ation event taking place, with the expected rates as set in the simulation. We find that the observed initiation rate is less than the potential rate for all rates greater than 0.08 s^{-1} (Fig. 3.14, p. 125). This deviation is due to occlusion of the start codon by ribosomes which have not yet advanced enough. By taking into account the combined time for both binding and clearing the start codon, we find that the observed initiation frequences more closely match the theoretical values (Fig. 3.15, p. 126). Yet, once initiation surpasses a rates of $\sim$0.66 s^{-1}, the observed initiation rate still reaches a plateau much sooner than predicted. Direct occlusion of the start codon by a ribosome is insufficient to explain the observed reduction in the observed initiation rate.

To better understand this effect, we simulated the "observed" *elongation* rates after changing the "expected" initiation rates by modifying the probability of loading a ribosome onto the start codon. The translation elongation rates are clearly affected by the initiation rates (Fig. 3.16, p. 128). An initiation rate of 0.04 s^{-1} yields the expected translation rate of 12.5 s^{-1}, while all other elongation rates decrease. This relationship plateaus at just over 10 a.a. s^{-1}, and the decrease is more pronounced at the beginning of the mRNA (Fig. 3.17, p. 129). Since higher initiation rates are able to more densely populate an mRNA with ribosomes, the translation rates decrease because of more frequent collisions between

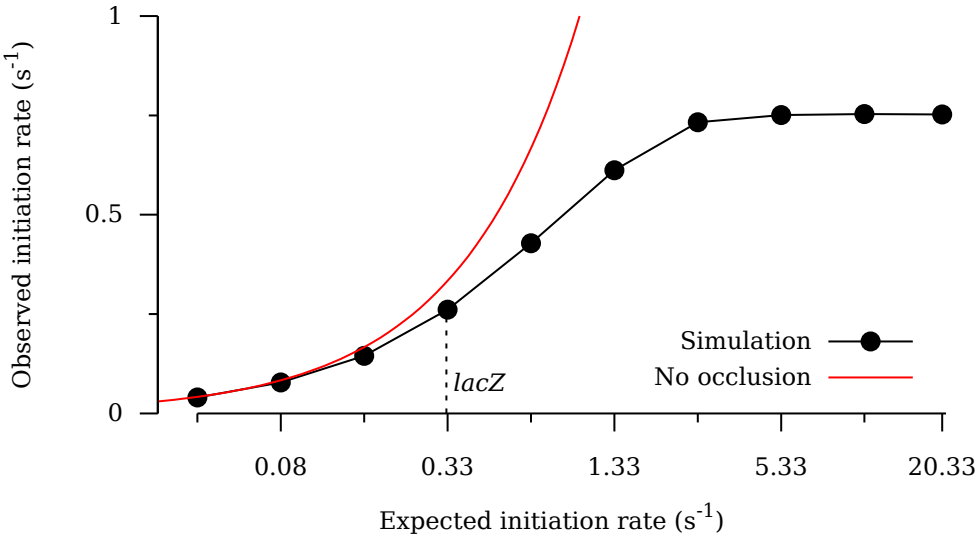


Figure 3.14: Occlusion of the start codon can prevent translation initiation. Different initiation rates were simulated 100 times each, and the observed rates were compared to the expected simulation parameters. Any decrease in the observed initiation rate is due to occlusion of the start codon by a ribosome which has not yet advanced enough. This occlusion is increasingly likely as initiation events become more frequent. Increasing the initiation rate parameter causes the observed rate to become progressively lower than the expected rate.

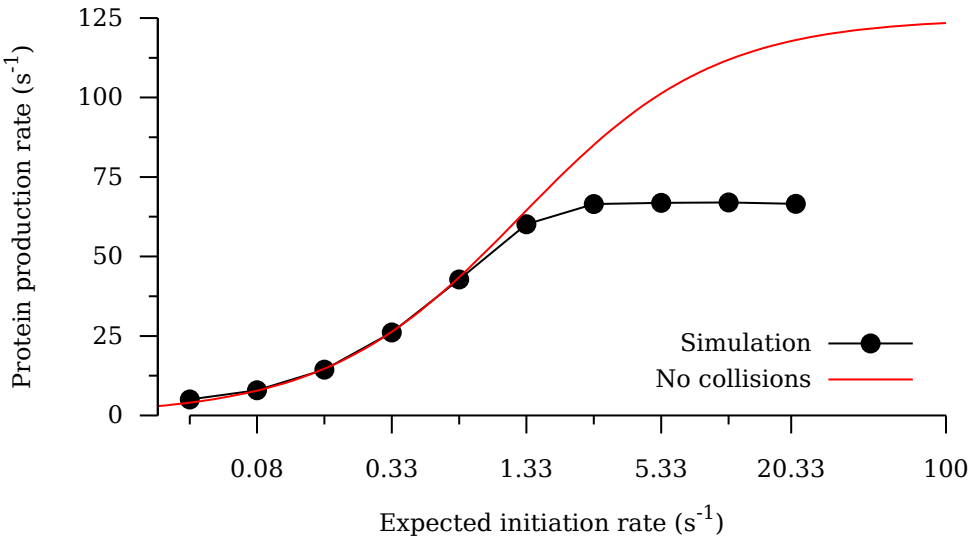


Figure 3.15: Ribosome collisions downstream of the start codon can cause occlusion of the start codon at very high initiation rates. For 100 mRNAs simulated with an average translation rate of 12.5 a.a. s^{-1}, the expected maximal protein production rate would be 125 s^{-1}, when initiation is instantaneous and ribosomes do not collide. In fact, the maximal production rate is much lower due to the formation of a queue throughout the mRNA, which ultimately blocks the start codon.

ribosomes. The increase in collisions in turn slows down translation near the start codon, which reduces the initiation rate. This behavior explains how our predicted protein production rates deviated from the simulation: a negative feedback loop between ribosomes blocking initiation and ribosomes colliding reduces both the initiation and elongation rates. Given a potential 12.5 a.a. s^{-1} elongation rate, this relationship results in a maximum observed initiation rate of roughly 0.75 s^{-1}.

We have shown how, in principle, ribosome queue formation can have a strong effect on protein production and on the average rate of ribosome movement. In general, we observe that the presence of slow codons can seed the formation of a ribosome queue, and have argued that ribosome profiling is well-suited to detect these types of ribosome pauses. In fact, it is often stated that ribosome profiles can be directly interpreted to directly test the correspondence between codon identity and the translation rate. Counter-intuitively, queue formation might actually explain a divergence between this relationship. If a small patch of slow codons causes a queue to form, then the ribosome density will increase at the upstream codons, regardless of whether those upstream codons are intrinsically fast or slow. Thus queuing effects might give rise to situations where both slow and fast codons have high ribosome occupancy. To test for the conditions under which fast codons can ac-

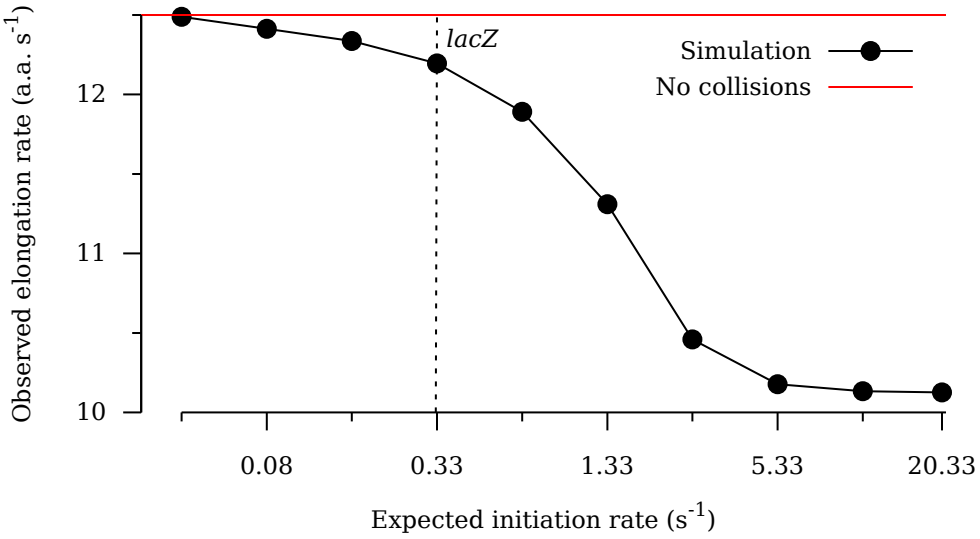


Figure 3.16: Average translation elongation rates decrease on densely packed mRNA. The average elongation rate over the mRNAs was computed by recording the average fraction of successful ribosome advances across the mRNA. As the translation initiation rate increases, and the spacing between ribosomes drops, the ribosomes are more likely to form queues. Therefore the translation elongation rates decrease. However, this effect is counter-balanced by blocking of the start codon, which in turn limits the maximum initiation rate. Here, the observed elongation rate of 12.5 a.a. s^{-1} drops to nearly 10 a.a. s^{-1} at very high initiation rates.

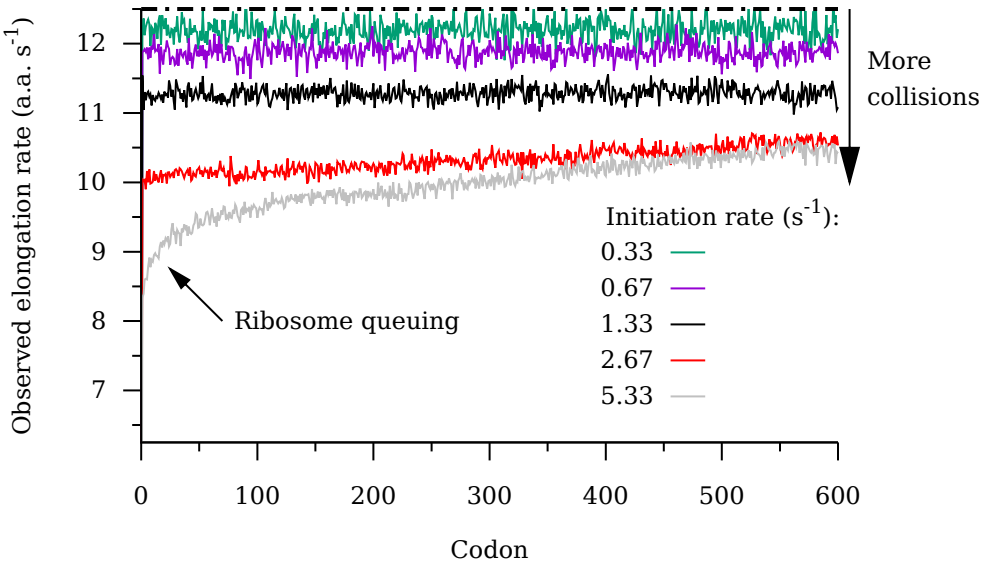


Figure 3.17: Average translation elongation rate decreases are more pronounced near the start codon. Shown are the first 600 observed translation rates for each codon, averaged across all ribosomes. All simulations had the same expected translation rate of 12.5 a.a. s^{-1} at each codon (dashed line), but the observed rates are always less. As the initiation rate increases, collisions between ribosomes are more frequent, causing translation to slow down. When the initiation rate is sufficiently high, these collisions disproportionately affect the codons immediately after downstream of the start codon. This slowdown ultimately sets a lower limit for the observed initiation rate.

cumulate high ribosome density, we simulated the translation of a short mRNA, 100 codons long, with alternating fast (15 s^{-1}) and slow (10 s^{-1}) codons. Next, we modulated the average ribosome density, and thus the probability for ribosomes to form a queue by varying the rate of translation initiation. As expected, under low initiation conditions we observe that ribosome occupancy of slow codons is higher than of fast codons (Fig. 3.18, p. 131). However, as the initiation reach reaches 1.33 s^{-1}, the ribosome occupancy of the upstream codons begins to increase, and when the initiation rate is 2.33 s^{-1} these fast codons actually have higher ribosome occupancy than the slow ones. It is therefore not necessarily the case that slow codons will always have higher ribosome occupancy than fast codons.

3.4 Discussion

Deciphering the individual contributions of codons to the translation rate along the mRNA is difficult. Ribosome profiling experiments provide nucleotide-resolution information on the relative ribosome densities, however interpreting these data in terms of translation rates has not yet been accomplished for several reasons. First, the density information does not directly report on translation rates, but rather the

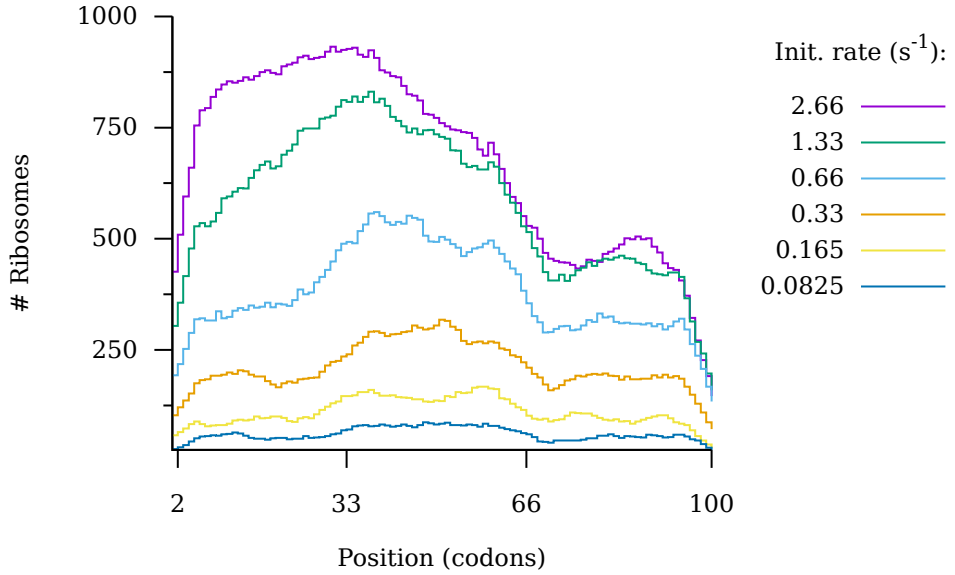


Figure 3.18: Ribosome queuing can increase ribosome occupancy of fast codons. An mRNA 100 codons long and containing alternating fast (15 s^{-1}, position 2-33), slow (10 s^{-1}, position 34-66) and fast (15 s^{-1}, position 67-100) codons was simulated for 240 sec. The simulation was performed 1,000 times, and the final ribosome occupancy of these 1,000 mRNAs was summed, yielding the total number of ribosomes occupying each codon. To determine the effects of the translation initiation rate, the simulation was performed with rates varying from 0.08625 s^{-1} to 2.66 s^{-1}. Ribosome occupancy of fast codons upstream of slow codons is observed to increase commensurately with the translation initiation rate.

relative probability for a ribosome to be located at a particular codon. Although it is the case that some codons are translated more rapidly than others, any ribosome queuing can produce the accumulation of ribosomes at any codons upstream of a pause site regardless of their intrinsic translation rates. Fast codons might sometimes have higher ribosome densities than slow codons. Predicting queue formation in turn requires knowing the translation initiation rate, which is unknown for most open reading frames. The theory of queue formation has been extensively studied, and understanding the propagation of ribosome queues whose formation depends on the local translation rates of hundreds or thousands of individual codons would not be trivial.

Second, it is known that high-throughput sequencing experiments have sequence-dependent biases which unevenly perturb the sequencing read number of different sequences (discussed in Ch. 2.4.2, p. 86). These biases are self-correcting either when comparing identical sequences, in which these biases are the same, or large sequence regions where the biases could average out. As a result, mRNA-Seq has been very good at comparing gene expression changes between conditions, as the sequences of the mRNAs in questions are the same and the sequencing depth is spread out over entire mRNAs. Similarly, using ribosome profiling to compare ribosome allocation to transcripts has also

been very effective ([Li et al., 2014], and *e.g.* Ch. 4.3.4, p. 176). However, quantifying the ribosome densities across an mRNA does not benefit from either of these self-correcting properties. As a result, the most convincing demonstrations using ribosome profiling that some codon types have higher occupancy than others have relied on calculating averages across many mRNAs.

We looked for a relationship between ribosome pauses, as measured using ribosome profiling, and the translation elongation rate by specifically studying *E. coli lacZ*, a well-characterized mRNA. In particular, the *lacZ* translation initiation and elongation rates have already been determined [Kennell and Riezman, 1977, Sørensen et al., 1989], and we have already collected ribosome profiling data from cells expressing the mRNA (Ch. 4, p. 136 and Ch. 5).

Clearly, some combinations of translation initiation and elongation rates promote ribosome collisions, slowing down translation. Our simulations predict that small amounts of ribosome queuing are possible on a *lacZ* mRNA is possible given fixed initiation and elongation rates. With these parameters, the mRNA is in a regime which balances the allocation of ribosomes to the mRNA with a limited number of collisions. In general, queue formation requires much higher initiation rates than have been observed, which might indicate that the initiation and elon-

gation rates have been coordinately optimized to balance efficient ribosome allocation against inefficient translation elongation. In fact, it has been suggested that slow codons at the beginnings of genes, combined with mRNA instability, can help reduce the formation of queues and promote efficient translation [Mitarai et al., 2008]. Since queuing appears to be limited for the *lacZ* mRNA, queue formation might not be sufficient to explain the large variations in ribosome density which we have measured using ribosome profiling. However, if some regions of the mRNA are translated more slowly, then even short stretches of "slow" mRNA might be enough to seed the local formation of ribosome queues. Yet previous experiments, in which inserts into the *lacZ* mRNA altered the translation rate of the protein, were not shown to promote ribosome queue formation, except for a single extremely slow insert [Sørensen and Pedersen, 1991]. Our ribosome profiling data might indicate the existence of such localized queues.

Interestingly, simulating the pulse-chase experiment suggests that the *lacZ* mRNA translation rate increases. Although the cause of this rate change is unknown, we have found that very slow translation at the beginning of an mRNA could regulate the initiation of ribosomes by occluding access to the mRNA. Using this information, we simulated the ribosome density along the *lacZ* mRNA. As expected, the transition be-

tween two translation rates produces a single drop in ribosome density at the position where the rate changes. Since the shape of the simulated density is less complex than our experimental data, the *lacZ* mRNA translation rate might be more variable than previously known. However, there is a clear discrepancy between the measured average translation rate and ribosome profiling along the *lacZ* mRNA, though the reason for this discrepancy is not known. It may be that current ribosome profiling protocols do not adequately capture the relative probabilities of ribosome pausing, or that simple collision models of ribosomes programmed with fixed translation rates are too simplistic. Ribosome profiling studies of mRNAs which are regulated by poor ribosome binding sites may prove to be particularly informative, as these mRNAs would be expected to have fewer ribosomes simultaneously occupying the mRNA, and thus the ribosomes would be less likely to collide or form queues.

Chapter 4

Tunability of ribososome density on translating mRNA

During translation multiple ribosomes sequentially engage an mRNA to form a polysome. Efficient translation requires rapid progression of ribosomes beyond the start codon and results in tighter ribosome packing. In bacteria, this process is regulated by ribosome binding upstream from the start codon and the translation of the first codons. Fast translation is facilitated by codons recognized by abundant tRNA recognized by Watson-Crick (WC) base-pairing. Although ribosome pauses at slow codons can affect protein production, the phenomenon remains poorly understood because translation rate measurements are notoriously difficult. Here we show, using ribosome profiling, that synony-

mous changes to an mRNA can predictably modulate both the translation of early codons after translation initiation, determining protein synthesis levels, and average translation rates, re-distributing ribosome density on a synonymously re-coded mRNA. We also find that the 15 codons immediately downstream of the start codon regulate Luc translation. The combination of low translation levels and slow translation elongation lead to increased abortive translation. These results demonstrate that synonymous codons can sometimes determine translation rates, affecting polysome spacing, and also influence protein production.

4.1 Introduction

There are 61 sense codons in the genetic code, conserved from bacteria to humans, yet no organism codes the full complement of tRNA anti-codons [Ninio, 1971, Chan and Lowe, 2009]. Codons lacking a tRNA whose anti-codon is perfectly complementary are recognized via wobble [Crick, 1966]. Transcriptomes contain both types of codons, and wobble translation is ubiquitous. When the genetic code was deduced, it was realized that some amino acids have many (up to six) codons, while others have just one [Khorana, 1968]. Much as two words which share

the same meaning are synonyms, these codons encode the same amino acid and were originally termed synonymous. This analogy implies that synonymous codons are functionally identical, however, several features differentiating synonymous codons have since been discovered, for example: providing binding sites [Thurman et al., 2012, Stergachis et al., 2013], facilitating mRNA secondary structure [Mita et al., 1988], regulation of gene expression [Sharp and Li, 1987], and modulation of translation rates [Sørensen and Pedersen, 1991]. These functions can overlap, with each contributing to the sequence composition of genes. Distinguishing between the different roles for synonymous codons has therefore been challenging.

The wild-type *E. coli* translation rate typically ranges from 10 to 20 a.a. sec^{-1} (Ch. 1.5, p. 24). This rate is reduced by proofreading [Ninio, 2006] and by wobble recognition [Sørensen and Pedersen, 1991]. It has also been suggested that greater tRNA abundance promotes optimal translation [Ikemura, 1981b]. Experiments in *E. coli* provided evidence that synonymous changes to the mRNA of Firefly Luciferase (Luc), a reporter gene, could predictably shift translation rates from intermediate to fast [Spencer et al., 2012]. These changes were facilitated *via* a two-part strategy using abundant tRNA and WC recognition. To address how synonymous codons affect translation we have analyzed the ribosome

densities of synonymously re-coded Luc mRNA (Fig. 4.3, p. 160) using ribosome profiling [Ingolia et al., 2009], a high-throughput sequencing method which measures the relative abundance of ribosome-protected mRNA fragments. In theory, ribosome profiling is suited to address this question because ribosome density is inversely proportional to the ribosomal transit time. We find that ribosome densities on the mRNA can be rationally modulated over long sequence stretches by introducing synonymous changes, although the correspondence between ribosome densities and translation efficiency is often poor over shorter sequence stretches.

4.2 Methods

4.2.1 Re-coding Luciferase mRNA using the translation index

Two principal factors have been suggested to cause differential translation rates of synonymous codons: the type of tRNA:mRNA recognition (WC *vs* wobble), and the cellular tRNA levels. The kinetics of translation are important for the accuracy of protein synthesis [Ehrenberg et al., 1986], which depends on the discrimination of different tRNAs

and anti-codons. Although our understanding of the kinetic impairment caused by wobble recognition is incomplete, it was demonstrated that the WC and wobble codon pair encoding Glu have a threefold different translation rate [Sørensen and Pedersen, 1991]. The role of cellular tRNA levels is more controversial, and both relative and absolute concentrations may be important. The relative levels may regulate the probability of finding a matching tRNA amidst all other possible tRNAs in the cell [Varenne et al., 1984], while high absolute tRNA levels may be necessary in order to saturate the ribosomes at high growth rates [Ehrenberg and Kurland, 1984]. Based on experimental measurements, it has been suggested that the gene copy number is a good proxy for cellular levels in many organisms [Ikemura, 1981b, Ikemura, 1981a, Ikemura, 1982, Ikemura and Ozeki, 1983, Ikemura, 1985, Dong et al., 1996]. However, there are many caveats which must be kept in mind when using tRNA gene copy numbers to estimate tRNA levels in the cell for the purpose of inferring translation rates. First, not all tRNAs are aminoacylated at all times, and the fraction of charged tRNA can decrease under conditions of amino acid starvation to varying degrees. Thus some abundant tRNAs might have lower populations of active molecules than others. Second, chromosomal gene copy number is a crude means of estimating expression level, as it is known that gene expression is com-

plex, depends on proximity to the origin of replication, and can be modulated in response to external conditions. Thus a low copy number gene could be highly transcribed, and *vice-versa*.

Starting from these two concepts, WC or wobble recognition and high or low tRNA concentration, many attempts have been made to rank codons based on their relative translation rates (see Ch. 1.2, p. 7). Since very few absolute codon-specific translation rates are known, it has been a challenge to predict to what extent the translation rates of different mRNA sequences diverge. In general, these rankings serve to develop a general expectation for which codons within a set of synonymous codons are, on average, translated more rapidly. These unitless scales cannot be readily converted into mean translation rates in codons translated per second, nor should they be considered as true translation rate predictors. Nevertheless, they serve as a good starting point from which different mRNA sequences can be compared or created *de-novo*, helping to develop a hypothesis for whether or not the translation rate will be affected by the mRNA sequence.

A translation index was previously developed to rank synonymous codons according to WC/wobble and tRNA gene copy number parameters [Spencer et al., 2012] (Table 4.1, p. 145). In brief, codons recognized by tRNAs with high copy number genes in the genome are

given higher weights than synonymous codons recognized by tRNAs with lower gene copy numbers. In addition, two synonymous codons recognized by the exact same tRNAs might have different weights, depending on whether recognition is WC or wobble; wobble recognition reduces the index for a codon by one third. This factor is derived from the observation that the translation of synonymous Glu codons in *E. coli*, which share the same tRNAs, were determined to differ by a factor of 3 [Sørensen and Pedersen, 1991]. Therefore, this index ranks codons using a combination of tRNA gene copy number, reflecting the delay for the diffusion of tRNA to the ribosome, and the type of recognition, reflecting delays which occur once a tRNA is being recognized. The index thus predicts which synonymous codons have a relatively higher or lower translation efficiency.

Luc mRNA variants were previously designed using this translation index, yielding a Fast Luc, exclusively composed of high-scoring codons, a Slow Luc, exclusively composed of low-scoring codons, and a Luc $\text{WT}_{\textit{E. coli}}$ which was re-coded so as to match the predicted WT firefly mRNA's translation efficiency. The unmodified wild-type mRNA (Luc $\text{WT}_{\text{Firefly}}$) was also analyzed. Translation indices were compared by looking up each value at the corresponding position within the Luc coding sequence. These values were then processed using window averaging

(window size of 31 codons). These alleles all end with a peroxisome-inactivation sequence, a linker, a c-myc tag and a His_6 tag, which were not re-coded.

Table 4.1: Translation index

Amino acid	Codon	WC Anticodon	Translation index
Ala	GCU	AGC	0.133
Ala	GCC	GGC	0.400
Ala	GCA	UGC	0.600
Ala	GCG	CGC	0.200
Gly	GGU	ACC	0.222
Gly	GGC	GCC	0.667
Gly	GGA	UCC	0.167
Gly	GGG	CCC	0.222
Pro	CCU	AGG	0.111
Pro	CCC	GGG	0.333
Pro	CCA	UGG	0.333
Pro	CCG	CGG	0.444
Thr	ACU	AGU	0.133
Thr	ACC	GGU	0.400
Thr	ACA	UGU	0.200
Thr	ACG	CGU	0.467
Val	GUU	AAC	0.095
Val	GUC	GAC	0.286
Val	GUA	UAC	0.714
Val	GUG	CAC	0.238

Ser	UCU	AGA	0.133
Ser	UCC	GGA	0.400
Ser	UCA	UGA	0.200
Ser	UCG	CGA	0.267
Ser	AGU	ACU	0.067
Ser	AGC	GCU	0.200
Arg	CGU	ACG	0.190
Arg	CGC	GCG	0.571
Arg	CGA	UCG	0.190
Arg	CGG	CCG	0.143
Arg	AGA	UCU	0.143
Arg	AGG	CCU	0.190
Leu	CUU	AAG	0.042
Leu	CUC	GAG	0.125
Leu	CUA	UAG	0.125
Leu	CUG	CAG	0.524
Leu	UUA	UAA	0.125
Leu	UUG	CAA	0.167
Phe	UUU	AAA	0.333
Phe	UUC	GAA	1.000
Asn	AAU	AUU	0.333
Asn	AAC	GUU	1.000
Lys	AAA	UUU	1.000
Lys	AAG	CUU	0.333
Asp	GAU	AUC	0.333

Asp	GAC	GUC	1.000
Glu	GAA	UUC	1.000
Glu	GAG	CUC	0.333
His	CAU	AUG	0.333
His	CAC	GUG	1.000
Gln	CAA	UUG	0.500
Gln	CAG	CUG	0.667
Ile	AUU	AAU	0.333
Ile	AUC	GAU	1.000
Ile	AUA	UAU	0.167
Met	AUG	CAU	0.500
Tyr	UAU	AUA	0.333
Tyr	UAC	GUA	1.000
Cys	UGU	ACA	0.333
Cys	UGC	GCA	1.000
Trp	UGG	CCA	1.000

Codons are highlighted according to the fast and slow coding strategies. "Fast" codons are indicated in green, while "slow" codons are indicated in orange.

4.2.2 Low-copy Luc vectors

To express Luc with high, robust levels of transcription, while simultaneously keeping total protein expression levels low enough to avoid

overexpression artifacts, Luc alleles were cloned into the low-copy number pACYC177 [Chang and Cohen, 1978] plasmid downstream of the P_{lac} promoter and ribosome binding site (Figure 4.1, p. 147). PCR products were derived from vectors bearing synonymously re-coded Luc genes, which were previously generated [Spencer et al., 2012], and from the *E. coli* MG1655 chromosome; these were combined using the Gibson assembly method [Gibson et al., 2009] (Table 4.3, p. 149). The Luc gene, including the transcriptional terminators from the pBAD vectors, was inserted on the strand opposite the one encoding the two antibiotic resistance genes to prevent transcriptional read-through. PCR products were prepared using the oligonucleotides listed in Table 4.2 (p. 148). Vector pACYC177 was digested with restriction enzyme BamHI-HF (New England Biolabs cat. no. R3136S), and the assemblies were performed as instructed by the kit (New England Biolabs cat. no. E2611).

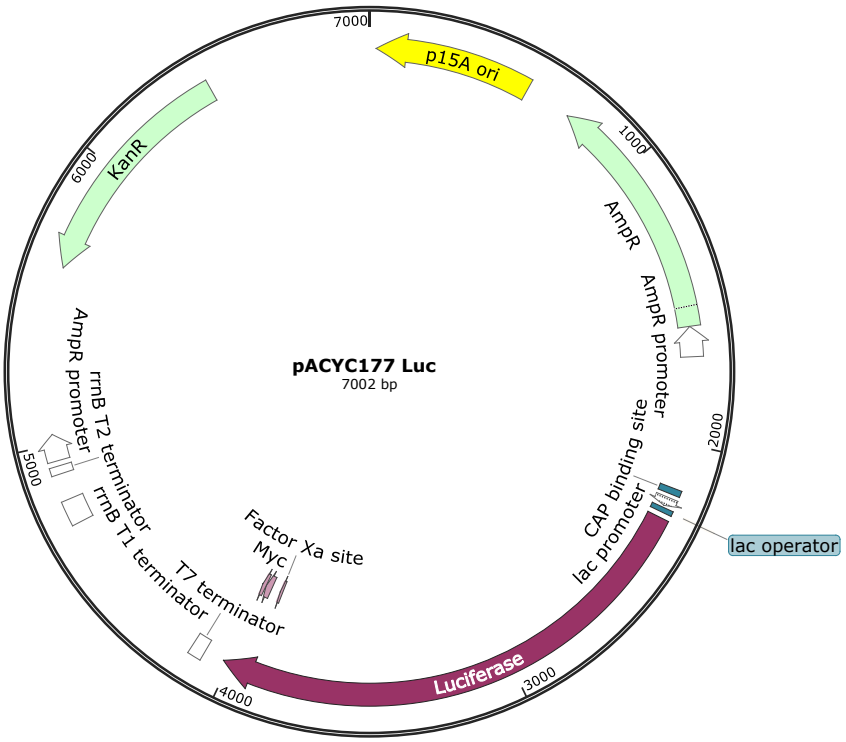


Figure 4.1: Low copy Luc plasmid construction. The Luc CDS (coding sequence) was cloned downstream of the *E. coli lac* operon promoter and ribosome binding site on the opposite strand from two antibiotic resistance genes, providing resistance to ampicillin and kanamycin.

Table 4.2: PCR Primers

ID	Description	Sequence
ATM13	P_{lac}-RBS Forward	CATTCGTGCGCTCTAGCCGCTTGCTGCAACTCTC
ATM14	P_{lac}-RBS Reverse 1 (F, WT, $WT_{Firefly}$)	GTCTTCCATAGCTGTTTCCTGTGTGAAATTG
ATM15	P_{lac}-RBS Reverse 2 (S)	GCATCCTCCATAGCTGTTTCCTGTGTGAAATTG
ATM16	Luc Forward 1 (F, $WT_{Firefly}$)	GAAACAGCTATGGAAGACGCAAAAAAC
ATM17	Luc Forward 2 (WT)	GAAACAGCTATGGAAGACGCTAAAAACATC
ATM18	Luc Forward 3 (S)	GAAACAGCTATGGAGGATGCTAAGAATATAAAG
ATM19	Luc Reverse	GGCTGAACGCCGGAGACTCTTCCTTTTTCAATATTATTGAAG
ATM33	P_{lac}-RBS Forward 2	CCCTCATTCGTGCGCTCTAGCCGCTTGCTGCAACTCTC
ATM34	P_{lac}-RBS Reverse 2	CATCCTCCATAGCTGTTTCCTGTGTGAAATTG
ATM35	S Front Forward	GGAAACAGCTATGGAGGATGCTAAGAATATAAAG
ATM36	S Front Reverse	GCGAAGAAGGAAAAAAGAGTAGGAACAAGAAG
ATM37	F Front Forward	GGAAACAGCTATGGAAGACGCAAAAAACATC
ATM38	F Front Reverse	GCAAAAAAAGAGAACAGGGTCGGTACCAG
ATM39	F Back Forward	ACTCTTTTTTCCTTCTTCGCAAAATCCAC
ATM40	S Back Forward	ACCCTGTTCTCTTTTTTTGCTAAGTCTACTCTTATAG
ATM41	F/S Back Reverse	GCACAGGCTGAACGCCGGAGACTCTTCCTTTTTCAATATTATTGAAG
ATM42	F16 Reverse	CATCCTCAAGCGGGTAGAACGGTGCCGG
ATM43	S569 Forward	GTTCTACCCGCTTGAGGATGGTACTGCTG
ATM44	S16 Reverse	TCTTCCAGAGGATAAAAAGGAGCAGGACC
ATM45	F569 Forward	TTTTTATCCTCTGGAAGACGGCACCGCA

Table 4.3: Gibson assemblies

Insert 1	Insert 2	Insert 3	Vector	Product
2	5	-	pACYC177	pACYC177/LucS
1	3	-	pACYC177	pACYC177/LucF
1	4	-	pACYC177	pACYC177/LucWT$_{\text{Firefly}}$
1	6	-	pACYC177	pACYC177/LucWT$_{E.\ coli}$
9	11	-	pACYC177	pACYC177/LucFS
7	8	10	pACYC177	pACYC177/LucSF
12	15	-	pACYC177	pACYC177/LucF$_{\text{in}}$S
13	14	-	pACYC177	pACYC177/LucS$_{\text{in}}$F

4.2.3 PCR

Amplification of DNA by PCR was performed using Q5 (New England Biolabs cat. no. M0491) or Phusion (New England Biolabs cat. no. E0553) polymerase (200 µM dNTP, 10 µM forward primer, 10 µM reverse primer, 1-50 ng DNA, 0.02 U/µl enzyme) in 50 µl reactions as directed by Table 4.4 (p. 150), using a Veriti 96-well thermocycler (Applied Biosystems). *E. coli* MG1655 chromosomal DNA was prepared using the GeneJet Genomic DNA Purification Kit (ThermoFisher cat. no. K0721).

Table 4.4: Polymerase chain reactions

DNA #	Forward Primer	Reverse Primer	Denaturation (initial)	Denaturation	Annealing	Extension	# Cycles	Template	Polymerase
1	ATM13	ATM14	98°C/30s	98°C/10s	64.4°C/25s	72°C/8s	35	*E. coli* MG1655	Q5
2	ATM13	ATM15	98°C/30s	98°C/10s	64.4°C/25s	72°C/8s	35	*E. coli* MG1655	Q5
3	ATM16	ATM19	98°C/30s	98°C/10s	59°C/25s	72°C/1m30s	35	pBAD18/LucF	Q5
4	ATM16	ATM19	98°C/30s	98°C/10s	59°C/25s	72°C/1m30s	35	pBAD18/LucW	Q5
5	ATM18	ATM19	98°C/30s	98°C/10s	59°C/25s	72°C/1m30s	35	pBAD18/LucS	Q5
6	ATM17	ATM19	98°C/30s	98°C/10s	59°C/30s	72°C/1m30s	35	pBAD18/lucH	Q5
7	ATM33	ATM34	98°C/30s	98°C/10s	64.4°C/25s	72°C/8s	35	*E. coli* MG1655	Q5
8	ATM35	ATM36	98°C/30s	98°C/10s	60.5°C/25s	72°C/50s	35	pACYC177/LucS	Q5
9	ATM33	ATM38	98°C/30s	98°C/10s	60°C/25s	72°C/50s	35	pACYC177/LucF	Q5
10	ATM39	ATM41	98°C/30s	98°C/10s	60°C/25s	72°C/50s	35	pACYC177/LucF	Q5
11	ATM40	ATM41	98°C/30s	98°C/10s	60°C/25s	72°C/50s	35	pACYC177/LucS	Q5
12	ATM13	ATM42	98°C/30s	98°C/5s	68°C/10s	72°C/1m8s	35	pACYC177/LucF	Q5
13	ATM13	ATM44	98°C/30s	98°C/5s	65°C/10s	72°C/9s	35	pACYC177/LucS	Q5
14	ATM45	ATM19	98°C/30s	98°C/5s	68°C/10s	72°C/1m8s	35	pACYC177/LucF	Q5
15	ATM43	ATM19	98°C/30s	98°C/10s	60°C/10s	72°C/1m30s	35	pACYC177/LucS	Phusion

4.2.4 Frameshift insertion

To insert two point mutations within the coding region of the Luc F gene attempts were made using conventional PCR-based “QuikChange” methods. However, the re-coded gene’s low sequence complexity, due to extreme re-use of the same codons, prevented the successful amplification of the desired PCR products. Therefore an alternative strategy was employed. A DNA gBlock Gene Fragment (5′-CTCCCACGCACGTGGACC

CGATCTTCGGCAACCAGATCATCCCGGACACCGCAATCCTGTCCGTAGTACCGTTCCACCACG

GCTTCGGCATGTTCACCACCCTGGGCTACCTGATCTGCGGCTTCCGTGTAGTACTGATGTACC

GTTTCGAAGAAGAACTGTTCCTGCGTTCCCTGCAGGACTACAAAATCCAGTCCGCACTGCTGG

TACCGACCCTGTTCTCCTTCTTCGCAAAATCCACCCTGATCGACAAATACGACCTGTCCAACC

TGCACGAAATCGCATCCGGCGGCGCACCGCTGTCCAAAGAAGTAGGCGAAGCAGTAGCAAAAC

GTTTCCACCTGCCGGGCATCCGTCAGGGCTACGGCCTGACCGAAACCACCTCCGCAATCCTGA

TCACCCCGGAAGGCGACGACAAACCGGGCGCAGTAGGCAAAGTAGTACCGTTCTTCGAAGCAA

AAGTAGTAGACCTGGACACCGGCAAAACCCTGGGCGTAAACCAGCGTGGCGAACTGTGCGTAC

GTGGCCCGATGATCATGTCCGGCTACGTAAACAACCCGGAAGCAACCAACGCACTGATCGACA

AAGACGGCTGGCTGCACTCCGGCGACATCGCATACTGGGACGAAGACGAACACTTCTTCATCG

TAGACCGTCTGAAATCCCTGATCAAATACAAAGGCTACCAGGTAGCACCGGCAGAACTGGAAT

CCATCCTGCTGCAGCACCCGAACATCTTCGACGCAGGCGTAGCAGGCCTGCCGGACGACGACG

CAGGCGAACTGCCGGCAGCAGTAGTAGTACTGGAACACGGCAAAACCATGACCGAAAAAGAAA

TCGTAGACTACGTAGCATCCCAGGTAACCACCGCAAAAAAACTGCGTGGCGGCGTAGTATTCG

TAGACGAAGTACCGAAAGGCCTGACCGGCAAACTGGACGCACTAAAATCCGTGAAATCCTGAT
CAAAGCAAAAAAAGGCGGCAAATCCAAACTGATCGAAGGCCGCGGATCTGGTACTAGTGGCGG
GTC-3′) was ordered from IDT DNA, digested with restriction enzymes PmlI (New England Biolabs cat. no. R0532S) and SpeI (New England Biolabs cat. no. R3133S) for 1 h at 37°C in NEB CutSmart Buffer, and ligated into digested pACYC177/LucF with T4 DNA ligase (New England Biolabs cat. no. M0202L) overnight at 15°C. Ligation products were transformed into *E. coli* NEB 5α cells (cat. no. C2987) *via* heat shock, and colonies were screened on ampicillin (100 µg/ml) and kanamycin (50 µg/ml) LB (Luria-Bertani) plates. After confirmation by sequencing, plasmid was transformed into the sequenced *E. coli* K-12 MG1655 laboratory strain (Yale CGSC # 6300; F-, λ-, rph-1). The sequence alignment shows little similarity between the in-frame and out-of-frame sequences (Fig. 4.2, p. 153).

4.2.5 Ribosome profiling

MOPS minimal medium [Neidhardt et al., 1974] plates (0.6% glycerol, 0.2% casamino acids, 100 µg/ml ampicillin) were streaked with *E. coli* MG1655 cells bearing one of the pACYC177-derived plasmids. Glycerol and casamino acids, rather than glucose, were added to minimize catabolite repression while avoiding amino acid starvation. 2.5

```
LucFast   MEDAKNIKKGPAPFYPLEDGTAGEQLHKAMKRYALVPGTIAFTDA   45
LucF_FR   MEDAKNIKKGPAPFYPLEDGTAGEQLHKAMKRYALVPGTIAFTDA   45

LucFast   HIEVNITYAEYFEMSVRLAEAMKRYGLNTNHRIVVCSENSLQFFM   90
LucF_FR   HIEVNITYAEYFEMSVRLAEAMKRYGLNTNHRIVVCSENSLQFFM   90

LucFast   PVLGALFIGVAVAPANDIYNERELLNSMNISQPTVVFVSKKGLQK  135
LucF_FR   PVLGALFIGVAVAPANDIYNERELLNSMNISQPTVVFVSKKGLQK  135

LucFast   ILNVQKKLPIIQKIIIMDSKTDYQGFQSMYTFVTSHLPPGFNEYD  180
LucF_FR   ILNVQKKLPIIQKIIIMDSKTDYQGFQSMYTFVTSHLPPGFNEYD  180

LucFast   FVPESFDRDKTIALIMNSSGSTGLPKGVALPHRTACVRFSHARDP  225
LucF_FR   FVPESFDRDKTIALIMNSSGSTGLPKGVALPHRTACVRFSHARGP  225

LucFast   IFGNQIIPDTAILSVVPFHHGFGMFTTLGYLICGFRVVLMYRFEE  270
LucF_FR   DLRQPDHPGHRNPVRSTVPPRLRHVHHPGLPDLRLPCSTDVPFRR  270

LucFast   ELFLRSLQDYKIQSALLVPTLFSFFAKSTLIDKYDLSNLHEIASG  315
LucF_FR   RTVPAFPAGLQNPVRTAGTDPVLLLRKIHPDRQIRPVQPARNRIR  315

LucFast   GAPLSKEVGEAVAKRFHLPGIRQGYGLTETTSAILITPEGDDKPG  360
LucF_FR   RRTAVQRSRRSSSKTFPPAGHPSGLRPDRNHLRNPDHPGRRRQTG  360

LucFast   AVGKVVPFFEAKVVDLDTGKTLGVNQRGELCVRGPMIMSGYVNNP  405
LucF_FR   RSRQSSTVLRSKSSRPGHRQNPGRKPAWRTVRTWPDDHVRLRKQP  405

LucFast   EATNALIDKDGWLHSGDIAYWDEDEHFFIVDRLKSLIKYKGYQVA  450
LucF_FR   GSNQRTDRQRRLAALRRHRILGRRRTLLHRRPSEIPDQIQRLPGS  450

LucFast   PAELESILLQHPNIFDAGVAGLPDDDAGELPAAVVVLEHGKTMTE  495
LucF_FR   TGRTGIHPAAAPEHLRRRRSRPAGRRRRRTAGSSSSTGTRQNHDR  495

LucFast   KEIVDYVASQVTTAKKLRGGVVFVDEVPKGLTGKLDARKIREILI  540
LucF_FR   KRNRRLRSIPGNHRKKTAWRRSIRRRSTERPDRQTGRTKIREILI  540

LucFast   KAKKGGKSKLIEGRGSGTSGGSGGSGRSEQKLISEEDLHHHHHH   584
LucF_FR   KAKKGGKSKLIEGRGSGTSGGSGGSGRSEQKLISEEDLHHHHHH   584
```

Figure 4.2: Out-of-frame residues are dissimilar compared to in-frame residues. A Needleman-Wunsch alignment [Needleman and Wunsch, 1970] of the Luc F and Luc F_{Fr} amino acid sequences shows little similarity. (top: WT Luciferase and c-terminal tag; bottom: with frameshift insertion). Identical residues are highlighted in blue.

ml of MOPS medium, with 1 mM IPTG, were inoculated from a plate and grown overnight, shaking at 37°C. 200 ml of the same medium, in pre-warmed 2 L flasks, were then inoculated by diluting the overnight cultures more than one thousandfold and set to shake at 37°C in an Innova 40 incubator shaker (New Brunswick Scientific) at 200 RPM (cell doubling time = 50 min). Once the cultures reached an OD_{600} of 0.40 - 0.65 (measured using a Shimadzu UV-1800 spectrophotometer; 1 OD_{600} = $6 \cdot 10^8 \frac{\text{cells}}{mL}$; saturated cultures can surpass an OD_{600} of 5.0) they were filtered through a pre-warmed 0.22 µm nitrocellulose filter paper and 90 mm glass filtration system (Kontes) with vacuum (Whatman cat. no. 7182-009); the filtration took about 1 min. The apparatus was rapidly disassembled, and the cells were immediately scraped from the filter paper, using a pre-warmed scoopula, and submerged into a 50 ml Falcon tube (Corning cat. no. 352070) filled with liquid nitrogen (LN2). The cells were then transferred to a mortar immersed in, and containing, LN2, and 600 µl of cell lysis buffer (10 mM $MgCl_2$, 100 mM NH_4Cl, 20 mM Tris-HCl pH 8.0, 0.1% Nonidet P40, 0.4% Triton X-100, 500 µg/ml chloramphenicol) were dripped in. The cells were pulverized with the chilled mortar and pestle, whereupon the LN2 was allowed to boil off. The remaining powder was scraped with a chilled spatula into a chilled 1.5 ml Eppendorf tube and stored at -80°C overnight.

Cell lysates were centrifuged 15 min at 14,000 RPM and 4°C, and the clarified supernatant was transferred to a 1.5 ml Eppendorf tube. Between 250 µl and 500 µl of supernatant were then treated with 2.0 µl of micrococcal nuclease (Life Technologies cat. no. EN0181) and 5 mM $CaCl_2$ and incubated 1 h while gently rotating at room temperature. The reactions were quenched using 6 mM EGTA and then transferred to 3.5 ml thickwall polycarbonate ultracentrifuge tubes (Beckman Coulter cat. no. 349622). The samples were underlaid with 2.0 ml of 1 M sucrose and centrifuged 6.5 h on a Beckman XL-80K ultracentrifuge at 55,000 RPM and 4°C in a SW55 rotor. Following removal of the liquid, the invisible pellets were re-suspended with 700 µl of Qiazol reagent, and RNA was extracted as directed by the miRNEasy kit (Qiagen cat. no. 217004).

Samples were denatured at 70°C and then subjected to electrophoresis on a 15% polyacrylamide TBE-urea gel 65 min at 200 V. The gel was then stained using SYBR gold and imaged with a blue light transilluminator. Gel slices were taken from the regions corresponding to lengths between 20 to 45 nt, as indicated by 10 bp ladder (Invitrogen cat. no. 10821015) and RNA size markers (5′-AUGUACACGGAGUCGAGCUCAACCCGCAACGCGA-(Phos)-3′ and 5′-AUGUACACGGAGUCGACCCAACGCGA-(Phos)-3′). The samples were then prepared as directed in Ingolia *et al.* [Ingolia et al., 2012], with modifications: ProtoScript II reverse transcriptase (New England

Biolabs cat no. M0368) was used for reverse transcription, and different biotinylated oligonucleotides were used for the removal of *E. coli* rRNA sequences [Li et al., 2014]. The library was sequenced on an Illumina HiSeq 2500 by the JHU Genetic Resources Core Facility. Each experiment was performed in duplicate or triplicate.

4.2.6 Sequencing data analysis

FASTQ files were processed in a number of steps, leading to the assignment of integer counts of footprints per gene. First, the reads were trimmed by removing the adapter sequence, using fastx_clipper 0.0.13.2. Next, rRNA and tRNA sequences were filtered using Bowtie2 2.2.9 [Langmead and Salzberg, 2012]. The remaining reads were mapped to the *E. coli* MG1655 genome (NC_00913.3) and the appropriate pACYC177 vector sequence using Bowtie2, followed by a procedure, written in Perl, which corrects sequencing errors and merges identical reads [Martens et al., 2015]. Using the list of proteins and their corresponding genomic map points from PTT files, each gene was scanned for mapped reads and the totals were recorded to a new file. Reads mapped to *tufA* and *tufB* were merged, and an additional scan was performed to assign reads to *ssrA*. Differential gene expression analysis was performed using DeSeq2 1.12.3 and R 3.3.1 [Love et al., 2014].

Table 4.5: Number of experimental replicates

Strain	Number
C	3
F	3
S	3
$WT_{Firefly}$	3
$WT_{\textit{E. coli}}$	3
F/S	3
S/F	3
F_{in}/S	2
S_{in}/F	2
F_{Fr}	2

Densities along the re-coded Luc mRNA were compared as follows. Position-specific footprint densities were assigned from normalized 30-nucleotide, 3′-aligned data, followed by base mean calculations from the replicates (Table 4.5, p. 157) using DeSeq2, and then window averaging (window size 151 nt) (Fig. 4.6, p. 166).

4.2.7 Spearman correlations

Spearman correlations between the ribosome profiles of the Luc F and Luc F_{fr} mRNAs were calculated using the cor.test "spearman" method, with a confidence level of 0.95, in R v. 3.3.3.

4.3 Results

4.3.1 Synonymous codons control ribosome density

To test whether fast translation is associated with low ribosome density, and slow translation with high ribosome density, a fast (Luc F) version of the Luciferase mRNA was designed using codons recognized by high copy number tRNA with maximal WC base-pairing. Likewise, a slow (Luc S) version was designed using low copy number tRNA and maximal wobble. The position-specific translation indices for Luc F and Luc S reveal significant differences between the fast and slow mRNAs (Fig. 4.3, p. 160). For comparison, the sequence was also modified to natively encode the wild-type translation parameters from the firefly into *E. coli* (Luc WT) (Fig. 4.3). To generate an expectation for the ri-

bosome density following these changes, we simulated the translation of mRNAs containing different combinations of fast and slow codons using a stochastic simulator with collisions (Ch. 3; Fig. 4.5, p. 164). We find that transitions between translation rate regimes will cause much greater changes in the ribosome densities than translation at a constant rate.

If codon composition influences the mRNA translation rate, it might be expected that queues would either be formed or released at the interfaces of regions of fast and slow (Fig. 4.4, p. 161). Regions coded for fast translation followed by slow translation should produce a characteristic pattern consisting of low density at the 5' end followed by high density at the 3' end (Fig. 4.4). This situation would be tantamount to creating an internal bottleneck. Conversely, a region coded for slow translation, followed by a region of fast translation, should result in high ribosome density at the 5' end and low density at the 3' end of the mRNA (Fig. 4.4), a situation representing the release of a bottleneck.

Our data show that re-coded Luc sequences produce distinct ribosome density patterns. The Luc F and Luc WT profiles along the mRNA are approximately uniform (Fig. 4.6, p. 166), consistent with more or less constant translation rates across the mRNA, with the exception of the 3′ end of Luc F, which shows higher densities. In contrast, the Luc S

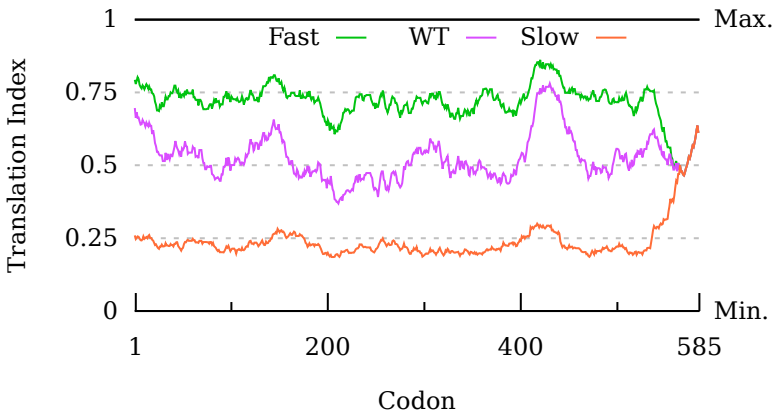


Figure 4.3: Synonymously re-coded Luc mRNA expressed in *E. coli* are predicted to have different translation rates. As expected, the translation index predicts that translation rates can be tuned within a threefold range by substituting synonymous codons with WC or wobble recognition.

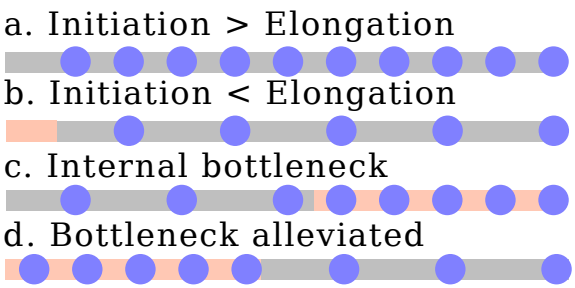


Figure 4.4: Polysome spacing is determined by the translation initiation and elongation rates. A polysome queue transitioning between fast (grey) and slow (red) translation induces changes in the local ribosome spacing. a. Ribosomes (blue circles) are more dense when the translation of the first codons promotes frequent initiation by new ribosomes. b. Ribosomes are less dense when initiate is less frequent. c. The transition from fast translation to slow translation produces an increase in ribosome density and a decrease in spacing. d. The transition from slow to fast translation results in decreased ribosome density.

allele shows comparatively low density near the start codon (Fig. 4.6). Both of these 5′ and 3′ density changes suggest that codon composition may regulate translation at the beginning and the end of the coding sequence (CDS), affecting translation initiation and termination. To induce the formation of queues the two halves of the Luc F and Luc S alleles were swapped, producing alleles Luc F/S and Luc S/F. True to expectations (Fig. 4.6c&e), we observe pronounced density changes. The fast-slow (Luc F/S) construct produces a bi-phasic pattern which starts out less dense and sharply transitions to greater density (Fig. 4.6c) and the slow-fast construct (Luc S/F) transitions from high ribosome density at the 5’ end to low density at the 3’ end (Fig. 4.6e).

To ascertain the contribution of codon composition to the translation of the initial codons, which might control translation initiation, two additional alleles were assembled. The first 15 codons (1.5 ribosome widths) downstream of the start codon were independently re-coded in the context of the Luc F and Luc S alleles, producing Luc F_{in}/S and Luc S_{in}/F, respectively (Fig. 4.6d&f). We find that these relatively small permutations produce large effects. Specifically, the 5′ end of Luc F_{in}/S has higher ribosome occupancy than Luc S, suggesting that translation initiation occurs more frequently and that the mRNA loads more efficiently. This result is supported by greater total translation of the Luc mRNA

(Fig. 4.7, p. 169). Opposite effects are apparent in cells expressing Luc S_{in}/F. High initial ribosome density at the 5′ end is followed by extensive low density (Fig. 4.6f). Translation levels of this mRNA are noticeably low, consistent with infrequent initiation (Fig. 4.7). Indeed, across all constructs, a clear pattern emerges whereby alleles that have fast coding in the first 15 codons downstream of the start codon have more than twofold the translation levels as the alleles with slow coding in the first 15 codons downstream of the start codon (Fig. 4.7). These results suggest that codon composition does affect initiation efficiency, as has been demonstrated before [Kudla et al., 2009, Pop et al., 2014, Evfratov et al., 2017], possibly by affecting the accessibility of the start codon to ribosomes which are not actively translating.

Figure 4.5: Polysome simulations of synonymously re-coded Luc mRNA predict large changes in ribosome density following the transition between translation rate regimes. Ribosome density profiles along the mRNA were simulated by changing the translation rates of different segments of the Luc mRNA, containing either fast codons (green) or slow codons (orange). The transition from fast to slow codons is predicted to cause an increase in ribosome density, while the transition from slow to fast codons causes a decrease. These mRNAs also encode the same c-terminal sequence with non-optimal codons (blue) as present in our experimental constructs (see methods).

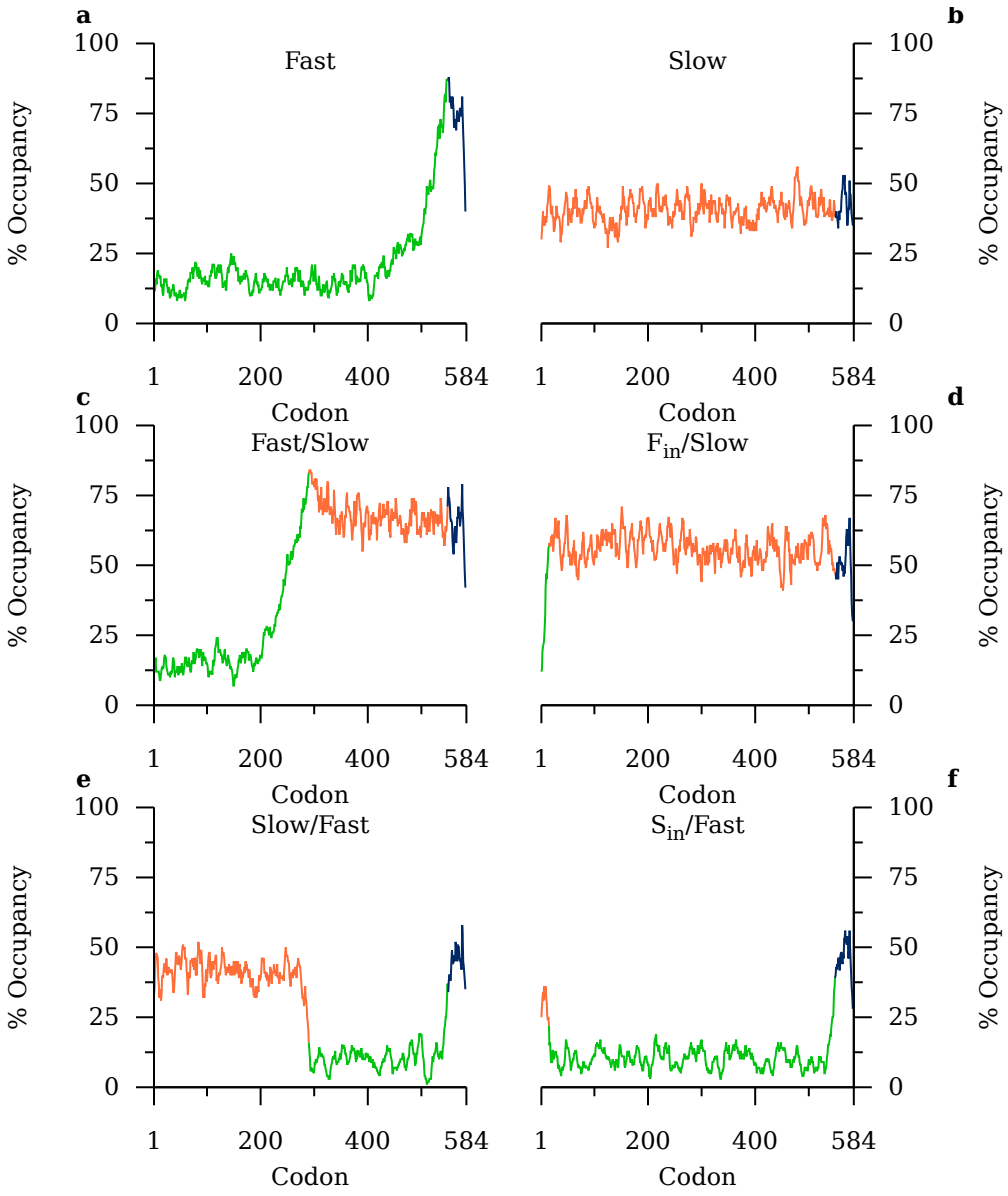
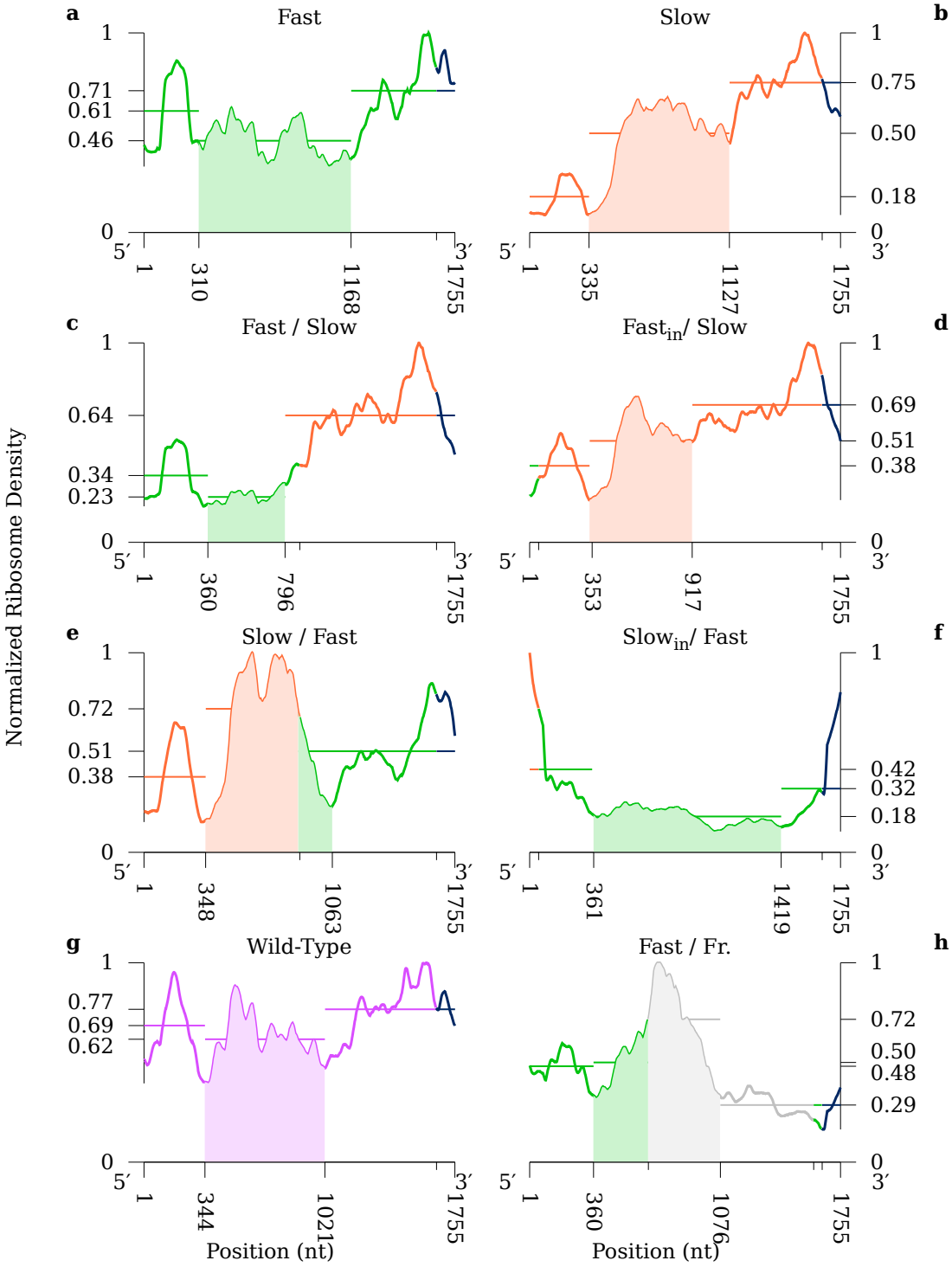

a
Fast
b
Slow
c
Codon
Fast/Slow
d
Codon
F_in/Slow
e
Codon
Slow/Fast
f
Codon
S_in/Fast
% Occupancy
100
75
50
25
0
1
200
400
584
Codon

Figure 4.6: Synonymously re-coding the Luc mRNA re-distributes ribosome densities. Using a re-coding strategy, the codons were changed to promote fast (green) or slow (orange) translation. A maximally occupied Luc mRNA is long enough to bind 62 ribosomes, assuming each ribosome occupies 28 nt and there is no intervening space between adjacent ribosomes.


a
Fast
1
0.71
0.61
0.46
0
5′
1
310
1168
1755
3′
b
Slow
1
0.75
0.50
0.18
0
5′
1
335
1127
1755
3′
c
Fast / Slow
1
0.64
0.34
0.23
0
5′
1
360
796
1755
3′
d
Fast$_{in}$/ Slow
1
0.69
0.51
0.38
0
5′
1
353
917
1755
3′
e
Slow / Fast
1
0.72
0.51
0.38
0
5′
1
348
1063
1755
3′
f
Slow$_{in}$/ Fast
1
0.42
0.32
0.18
0
5′
1
361
1419
1755
3′
g
Wild-Type
1
0.77
0.69
0.62
0
5′
1
344
1021
1755
3′
Position (nt)
h
Fast / Fr.
1
0.72
0.50
0.48
0.29
0
5′
1
360
1076
1755
3′
Position (nt)
Normalized Ribosome Density

We observe that the first ∼330 nt (∼11 ribosomes) and last ∼200 nt (∼7 ribosomes) of mRNA consistently yield an increase of ribosome density, regardless of the codons. The ribosome density (shaded color) following the initial sequence depends on the codon composition. Density is low when translation is faster (a, c), especially when total ribosome allocation levels are low (f). Density is high when translation is slower (b, d, e). The density increases abruptly when transitioning from fast to slow (c), and drops when transitioning from slow to fast (e). Density also increases with the wild-type coding strategy (g). The introduction of a frameshift produces a dramatic ribosome density peak (gray), which then decays (h).

4.3.2 Low ribosome densities are associated with abortive translation

E. coli mRNAs are degraded in the cell soon after their synthesis, with an average half-life of on the order of minutes [Forchhammer et al., 1972, Blundell et al., 1972, Pedersen et al., 1978, Chen et al., 2015], but actively translating ribosomes increase mRNA stability by occluding access of endonucleases to the mRNA [Morse and Yanofsky, 1969, Iost and Dreyfus, 1995]. Conversely, patches of naked mRNA are susceptible to

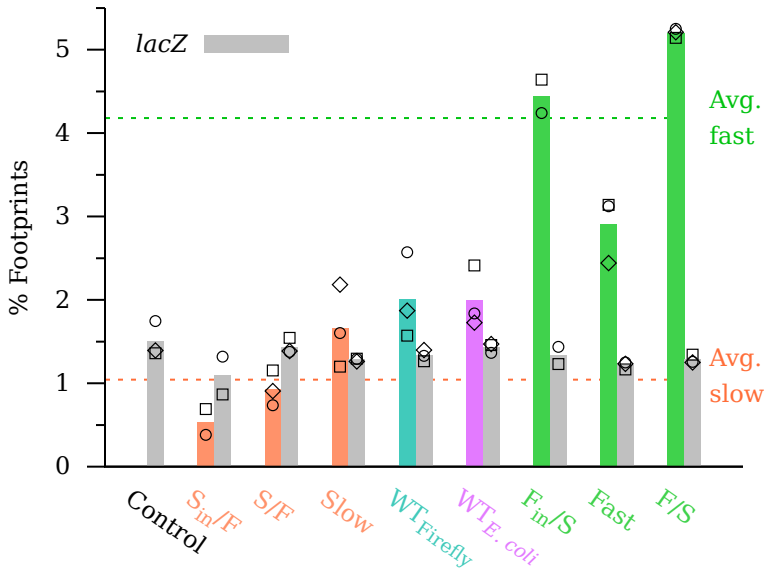


Figure 4.7: Synonymously re-coding the Luc mRNA affects ribosome allocation. The three mRNAs beginning with the slow codons (orange) have the lowest Luc translation levels. The three alleles beginning with the fast codons (green) have the highest translation levels. Shown also are the ribosome allocations of β-galactosidase (*lacZ*) (grey), which is also induced by IPTG and regulated by the same promoter and ribosome binding site sequences. Significant differences are found in the Luc levels, but no difference is seen in the *lacZ* levels.

nuclease attack, and ribosomes translating a partially degraded mRNA become stalled in the absence of a stop codon. Stalled translation on partially degraded mRNA can be rescued by trans-translation using a specialized tmRNA (reviewed by [Keiler, 2015]). The tmRNA inserts itself inside the ribosome, replaces the mRNA, and allows translation to resume by directing the addition of a short peptide sequence which encodes a C-terminal degradation tag.

Because active tmRNA (*ssrA*) is located inside the ribosome, ribosome profiling also quantifies these rescue events. We find that the expression of the Luc alleles does not affect the *ssrA* read number, except for allele S_{in}/F, which has double the *ssrA* abundance, indicating greater abortive translation in the cell (Fig. 4.8, p. 171). We conclude that the expression of a Luc mRNA which has low ribosome density stimulates the *ssrA* response.

4.3.3 Events other than ribosome pauses contribute to ribosome profiles

A major unresolved question is to what extent each codon, or group of codons, regulates the ribosome transit time along the mRNA. In principle, ribosome profiling has the necessary resolution to solve this issue.

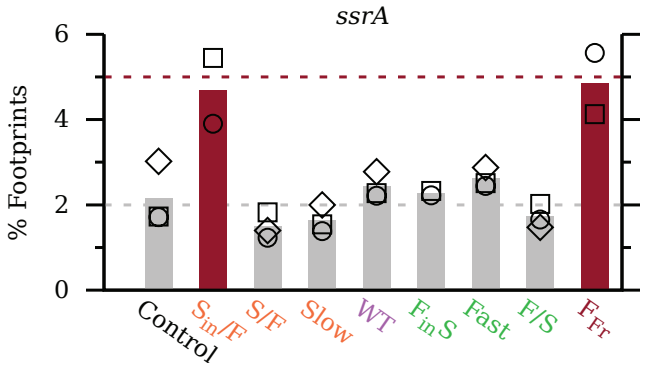


Figure 4.8: Low ribosome density is associated with increased ribosomal allocation to the tmRNA ribosome rescue system (*ssrA*). Infrequent translation initiation, combined with rapid translation elongation, leads to low ribosome density on the Luc S_{in}/F mRNA, possibly exposing the mRNA to nucleases, thereby inducing the formation of truncated Luc mRNA. Most Luc alleles have no effect on levels of tmRNA protected by ribosomes (~2-3%, grey). Expression of Luc S_{in}/F and, surprisingly, Luc F_{Fr} induces double the amount of ribosome-protected tmRNA (red).

However, ribosome profiling data are complicated by library preparation and sequencing biases [Artieri and Fraser, 2014a, Martens et al., 2015]. To evaluate these distortions we generated a new Luc variant, Luc F_{Fr}, which contains an internal frameshift spanning 311 codons yet only differs from Luc F by two nucleotides. The translation efficiency of the codons within this out-of-frame region is predicted to be lower than that of Luc F (Fig. 4.9, p. 173). Therefore, if the local ribosome profiles along the Luc mRNA depends on translation rates, then the profiles within this region will not be similar. However, if the mRNA sequence affects ribosome density independently of the translation rate, then they might be similar. We compared the Luc F and Luc F_{Fr} profiles without window averaging, to emphasize the shapes of the profiles at shorter ranges (Fig. 4.10, p. 174). As expected, the regions prior to the frameshift, which have identical sequences, have similar, overlapping peaks (Fig. 4.10). Prior to nucleotide #670 (the point where the reading frame is altered), we observe correlated peaks approximately centered at nucleotide positions 180, 240, 345, 465, 550 and 605. However, downstream of the frameshift insertion we observe ribosome stalling which is absent in Luc F. Downstream of this stalling point, the Luc F_{Fr} profile then gradually declines, while the Luc F profile does not. Thus at first glance, the insertion of a frameshift mutation perturbs the lo-

cal ribosome profile, as expected. Yet despite these major differences, the local fluctuations in the two signals are actually significantly similar (Fig. 4.11, p. 175). Downstream of nucleotide #840, we again observe correlated peaks, approximately centered at nucleotide positions 900, 945, 1020, 1065, 1185, 1290, 1405, 1505 and 1550, despite the difference in reading frame and predicted translation efficiency. Because these local similarities persist we conclude that contributions other than ribosome pauses significantly shape these profiles.

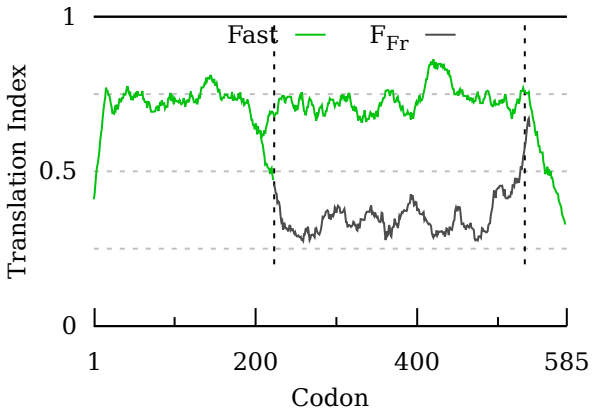


Figure 4.9: The frameshift region within Luc F_{Fr} is predicted to be translated more slowly than the Luc F sequence. The mRNA sequences differ only at two positions, but the amino acid sequence and codon composition within the frameshift are very different. Under the assumption that the translation rate depends on the mRNA sequence, we would not predict out-of-frame translation rates to be similar to in-frame translation rates.

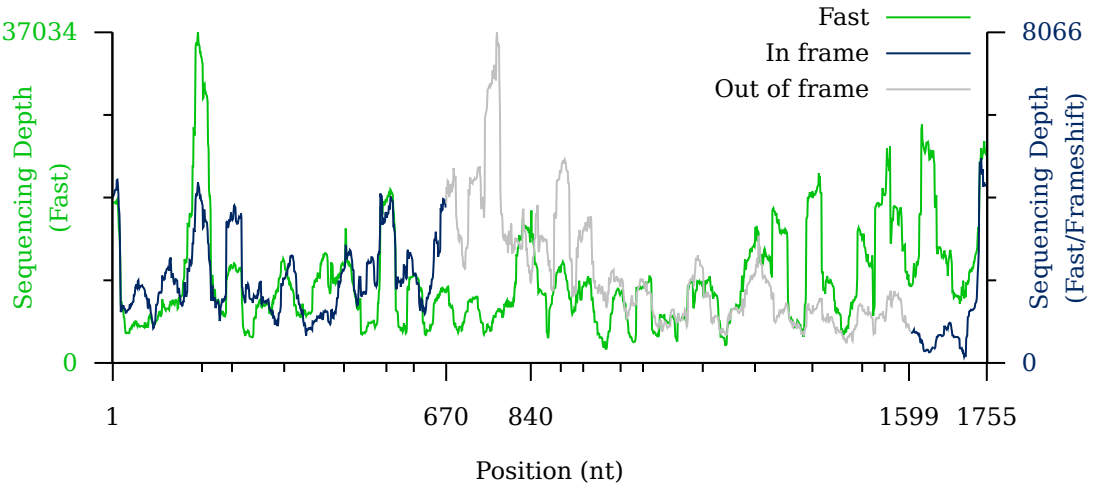


Figure 4.10: Local ribosome profiles are related to the mRNA sequence, not the predicted translation rate. A single base-pair was inserted into the Luc gene, producing a frameshift at position 670, which was then reverted by removing a base-pair at position 1599. The sequence between these two points (gray) encodes missense amino acids, but the mRNA sequence is identical. Local fluctuations in the magnitude of the ribosome profile clearly indicate the presence of contributions to the signal other than ribosome occupancy (Spearman correlation (ρ) = 0.31, $p < 0.05$).

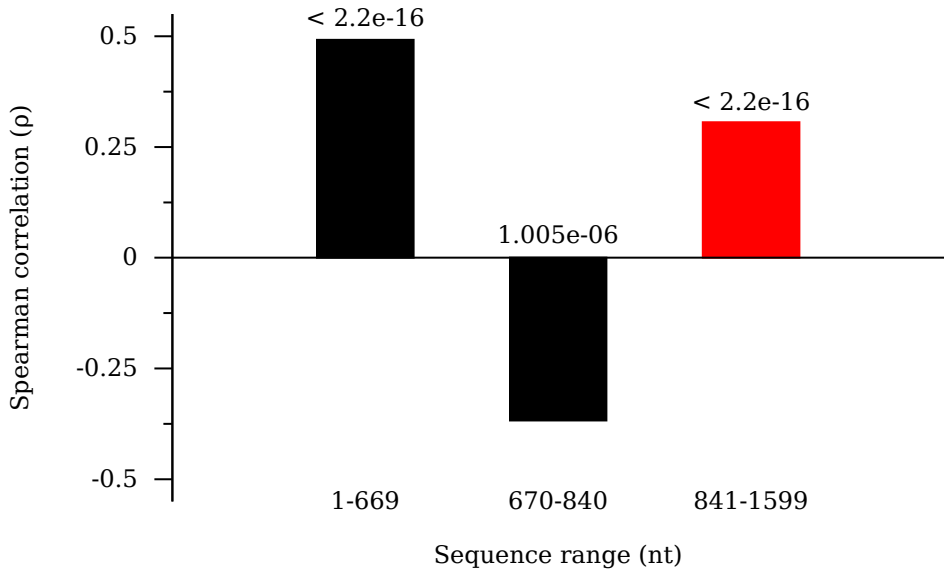


Figure 4.11: Out-of frame ribosome profiles are significantly correlated. Translation of 253 codons downstream of the frameshift produces a highly similar ribosome density profile (red), even though the codons and predicted translation rates are different. The ribosome profile upstream of the insertion point is highly similar, as expected, while the profile of the region immediately downstream of the insertion point is anti-correlated, possibly the result of ribosome stalling in the altered mRNA. The p-values are indicated above the bars.

4.3.4 Luc expression does not perturb ribosome synthesis

The extreme over-expression of mRNA bearing “non-optimal” codons leads to poor growth and ribosome degradation [Dong et al., 1995]. To determine whether expression of synonymously re-coded Luc affects the production of ribosomes we compared the fraction of ribosomes allocated to ribosomal protein production. We find that the the 54 ribosomal proteins command virtually identical fractions of the total number of ribosomes across experiments, indicating that ribosome synthesis is unaffected by Luc expression (Fig. 4.12, p. 177). Taken together, these genes comprise 14% of the total (doubling time of 50 min), in close agreement with the previously reported 15% (doubling time also 50 min) [Schleif, 1967].

4.4 Discussion

The polysome spacing is determined by relative, rather than absolute, translation rate differences. A useful analogy is that of cars traversing a single-lane toll bridge [Itano, 1968]. The time spent entering and leaving the bridge regulates the overall loading level, while the intro-

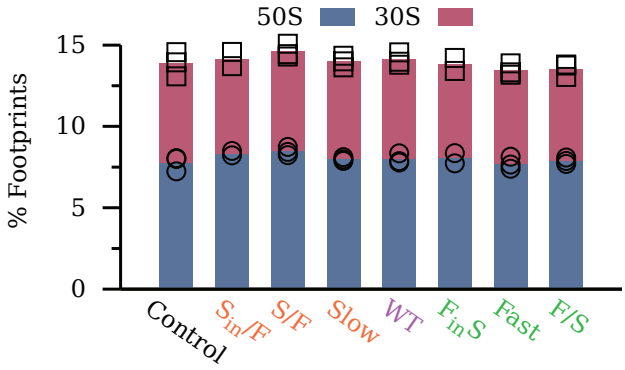


Figure 4.12: Ribosome allocation to ribosomal synthesis is unperturbed by the synonymous re-coding of the Luc mRNA. The total proportion of ribosome footprints allocated to the ribosomal proteins which comprise the large (50S, blue) or small (30S, red) subunit were summed, which together provide a measurement of the fraction of ribosomes allocated to ribosome production. This value (α_r), which depends on the growth rate, was previously reported to be 15%, when the doubling time was 50 min. [Schleif, 1967]. Our measurements ($\sim$ 14%) nearly match this value, when the doubling time was also 50 min.

duction of slow or fast points within the road creates and relieves traffic jams (Fig. 4.4). Previous ribosome profiling studies have shown evidence for the roles codons play in modulating translation rates, but technical challenges have made it difficult to quantify the effects in a given mRNA [Artieri and Fraser, 2014a, Gritsenko et al., 2015, Martens et al., 2015, Pop et al., 2014, Stadler and Fire, 2011, Woolstenhulme et al., 2015]. We systematically changed codons within the Luc mRNA to re-distribute ribosome density, and find that fast and slow codons have the expected effects, which are in agreement with extensive simulations [von der Haar, 2012]. However, these density changes are most apparent when explicitly forcing large translation rate changes over long sequence stretches. As evidenced from the comparison between the fast and frameshifted profiles, analyzed at shorter distances, current ribosome profiling protocols do not necessarily convey ribosome densities quantitatively. The faithful quantification of ribosome pauses with ribosome profiling remains unachieved.

Our data also quantitatively agree with previous translation rate measurements. Translation of the firefly Luc mRNA was measured to be half as fast as the Luc F mRNA (9.8 vs 19.2 a.a. sec^{-1}) [Spencer et al., 2012]. We find that fast codons are translated between two and three times faster than slow codons. Likewise, the transition from the initi-

ation phase to the elongation phase entails a greater drop in ribosome density when elongation is fast, and we observe more ribosome queuing when fast codons are near the 3′ end. The data are also consistent with altered mRNA stability. When less Luc protein is produced, as expected when translation initiation is infrequent, combined with fast translation, we detect abortive translation.

Comparisons of ribosome allocation to Luc mRNA also indicate that slow codons at the 5′ end are associated with low translation levels. The three Luc alleles which begin with slow codons have low total ribosome occupancy, whereas the three alleles which begin with fast codons have high occupancy. Interestingly, mRNA structure near the ribosome binding site has also been implicated in regulating protein expression [Kudla et al., 2009, Pop et al., 2014, Evfratov et al., 2017], and these two mechanisms might overlap.

Together, our data are consistent with less ribosome density when ribosomes translate faster codons, and greater ribosome density when fast ribosomes translate slower codons. We conclude that *E. coli* cells can simultaneously encode the same amino acid sequences and variable polysome density within the same mRNA sequence, which serves as an additional determinant of gene expression.

Chapter 5

The “heat shock” response to synonymously re-coded mRNA

The over-expression of heterologous proteins is an important tool in biotechnology. The conservation of the genetic code, the ribosome and chaperones make possible the expression of proteins in distantly related organisms. Yet there are many examples when such experiments either fail to produce protein, or the protein is sequestered into inclusion bodies [Shively, 1974]. It has been suggested that re-coding a foreign mRNA to more closely match the endogenous coding strategies of the cell can improve protein expression by synchronizing the co-translational protein folding with the translation elongation rate [Purvis et al., 1987, Krasheninnikov et al., 1989]. Here we show, using ribosome

profiling, that changes to the codons in the Firefly Luciferase mRNA affect the induction of chaperones, indicating changes in protein folding stress. However, after accounting for expression changes, we do not see any effects resulting from the translation of synonymous codons. Rather, these results demonstrate the serious burden of high protein expression on cells which are unable to sufficiently up-regulate chaperone production, with repercussions for the faithful production of folded protein.

5.1 Introduction

The maintenance of protein folding homeostasis is a serious challenge for the cell. Molecular chaperones, such as the “heat shock proteins” (HSPs), help to unfold and refold proteins, either co- or post-translationally, and are conserved from bacteria to mammals [Kim et al., 2013]. Chaperones form large protein complexes which consume ATP, therefore their synthesis is regulated to conserve resources and to promote growth. Bacteria, such as *E. coli*, regulate the expression of chaperones using a sigma factor (σ^{32}, in the case of *E. coli*), a transcription initiation factor which binds to RNA polymerase [Guisbert et al., 2008]. σ^{32} is unstable, has a low copy number and is kept inactive by

the chaperone complex DnaK/DnaJ (Hsp70/Hsp40). DnaK/DnaJ binds to unfolded proteins in response to folding stress, releasing σ^{32}. Free σ^{32} then associates with RNA polymerase and remodels the cell's transcriptional output, increasing chaperone production. The synthesis level of chaperones is therefore a direct readout of σ^{32} bound to DnaK/DnaJ, and of the levels of unfolded protein in the cell [Connolly et al., 1999]. Importantly, since σ^{32} activity directly responds to levels of unfolded protein, this “heat shock” response can be activated even in the absence of a temperature change.

The expression of eukaryotic proteins in bacteria often results in poor folding yields and protein aggregation, despite the presence of chaperones [Kim et al., 2009]. One potential cause is an incompatibility between eukaryotic and bacterial translation systems [Kurland and Gallant, 1996]. Although both the genetic code and translational mechanisms are highly conserved, organisms have evolved divergent sets of tRNA anti-codons [Chan and Lowe, 2009], and no organism has all 61 anti-codons [Ninio, 1971]. Intriguingly, genomes have evolved codon compositions which are “adapted” to these tRNAs [Grantham et al., 1980, Ikemura, 1981b, Ikemura, 1981a]. The mutual adaptation between mRNAs and tRNAs has been clearly demonstrated in specialized tissues which predominantly express a single protein: *Bombyx mori*

silk glands producing fibroin [Garel et al., 1970a], rabbit reticulocytes producing globin [Hatfield et al., 1982] and bovine eye lens producing crystallin [Garel et al., 1970b]. A similar adaptive mechanism may conceivably have evolved in the firefly *Photinus pyralis* light organ, where the bioluminescent Luciferase (Luc) protein is highly expressed.

The co-adaptation of mRNA sequences to the available tRNAs has been explained in terms of the ribosome’s kinetic proofreading mechanism. Translation of a limited codon repertoire can promote more efficient use of ribosomes and cellular resources, while promoting fast and accurate translation [Ehrenberg et al., 1986], and by reducing misreading errors resulting from tRNA competition [Kramer and Farabaugh, 2007, Kramer et al., 2010]. However, there may be instances where the presence of alternative codons could be advantageous. For example, co-translational folding of nascent proteins could be enhanced by specific mRNA sequences which synchronize the rate of emergence of the nascent polypeptide with the formation of new structural elements, improving co-translational folding [Purvis et al., 1987, Krasheninnikov et al., 1989]. Depending on the protein’s native structure, this synchronization may require either fast or slow translation. In contrast to evolutionarily adapted genes, genes transferred between distantly related organisms will probably not be well-adapted. Re-coding foreign mRNAs

to be compatible with the translation system of a new host could improve the folding of foreign proteins [Buhr et al., 2016].

Besides re-coding mRNAs, increasing the expression of chaperones has also been used to strategically improve the yield of folded protein when overexpressed. For example, it was shown that the overexpression of Trigger Factor can ameliorate the production of mammalian proteins in *E. coli* by reducing their aggregation [Nishihara et al., 2000]. Likewise, a two-step procedure which first co-induces chaperone expression, and then suppresses additional recombinant protein overexpression, was shown to improve solubility of the recombinant protein [de Marco et al., 2007]. Similar procedures have met with success, even for special glycoproteins [Skretas et al., 2009] or membrane proteins [Link et al., 2008], which are notoriously difficult to express in bacterial cells. As a result, the development of bacterial strains which are optimized for the production of foreign proteins has been of great interest [Makino et al., 2011].

To investigate the relationship between mRNA sequence and protein folding stress we used ribosome profiling [Ingolia et al., 2009] to monitor the induction of the σ^{32} “heat shock” regulon. We compared cells expressing different versions of synonymously re-coded Luc, a two-domain protein which requires chaperones to fold efficiently [Agashe

et al., 2004]. Surprisingly, we find that chaperone induction depends on the total levels of Luc protein in the cell, rather than the coding of the Luc mRNA *per se*. By measuring the effects of Luc protein accumulation, we develop a model which predicts that the σ^{32} response saturates following just modest levels of Luc expression. This result suggests that *E. coli*’s ability to manage protein folding stress is intrinsically limited. This limit is a major obstacle for the over-expression of proteins which require chaperones to fold.

5.2 Methods

5.2.1 Ribosome profiling

Ribosome profiling was performed as in chapter 4.2.5, p. 152.

5.2.2 Western blots

2.5 mL of MOPS minimal medium [Neidhardt et al., 1974] were inoculated with an *E. coli* colony from from a MOPS minimal medium plate and grown overnight, shaking at 37°C. Fresh 2.5 mL cultures were inoculated the following day to an OD_{600} of 0.01 and grown 6 hours (∼7 doublings) to an OD_{600} of ∼1.0 and then placed on ice. 1 mL aliquots were

transferred to chilled 1.5 mL Eppendorf tubes and centrifuged 5 min at 14,000 RPM at 4°C. Using the OD_{600} readings of the remaining cultures, the cellular pellets were resuspended in Laemmli sample buffer (50 µl / OD_{600}), and the cellular suspensions were boiled 10 min at 100°C and centrifuged 2 min at 14,000 RPM at room temperature. 2.5 µl Spectra BR protein ladder (ThermoFisher cat. no. 26634) and 7.5 µl cell lysate aliquots were subjected to electrophoresis by 10% SDS PAGE 50 V for 15 min, then 100 V for 75 min. Transfer was executed using Immun-Blot LF PVDF membrane (Bio-Rad cat. no. 162-0261). Blotting was performed with mouse anti-myc primary antibody (Cell-Signaling Technologies cat. no. 2276) and goat anti-mouse alexa-fluor 594 secondary antibody (Cell-Signaling Technologies cat. no. 8890) for Luc detection, and rabbit anti-β-galactosidase primary antibody (ThermoFisher cat. no. A-11132) and goat anti-rabbit alexa-fluor 488 secondary antibody (Life Technologies cat. no. A-11034). Primary blotting was performed overnight at 4°C and secondary blotting 30 min at room temperature. Imaging was performed using a ProteinSimple FluorChem M imager and processed using ImageJ v. 1.50h.

5.2.3 Enzyme assays

Luciferase activity was measured using the Dual-Light kit (ThermoFisher cat. no. T1004) with a Berthold TriStar2 plate-reader luminometer as follows. 2.5 mL MOPS minimal medium (0.6% glycerol, 0.2% casamino acids, 100 µg/ml ampicillin) cultures [Neidhardt et al., 1974] were prepared from a minimal medium plate and grown overnight. These were then diluted into 2.0 mL of medium to an OD_{600} of 0.01 and grown to and OD_{600} of 1.0 at 37°C. The cultures tubes were then immediately placed on ice, and for each culture, three 100 µl aliquots of medium containing cells were transferred to separate 1.5 ml Eppendorf microcentrifuge tubes containing 100 µl of Tropix lysis solution. Cells were lysed by vortexing for 10 sec. 10 µl of lysate were added to 25 µl of Buffer A (with luciferin) in a non-treated Nunc F96 MicroWell white polystyrene 96-well plate (ThermoFisher cat. no. 236105). After a single, initial 10 sec shake step (diameter 1), Luciferase activity was sequentially stimulated by injection of 100 µl of Buffer B (speed 5), mixed for 0.2 sec (diameter 5), and then, after a 2 sec delay, measured continuously for 15 sec with 0.05 sec intervals. To reduce background signal, no galacton+ substrate was added to Buffer B. Measurements were performed at room temperature (23°C). The luminescence was then averaged over the 15 sec duration, and the standard error of the mean was calculated for the

replicates.

5.2.4 Sequencing data analysis

Differential gene expression was analyzed using DeSeq2 [Love et al., 2014]. Integer numbers of sequencing reads, 1 per footprint, were assigned to either chromosomal *E. coli* genes or to genes encoded on the low-copy pACYC177 vectors bearing *bla*, *ampR*, *kanR* and *luc* genes (Fig. 4.1, p. 147). Since experiments on strains pACYC177, Luc F, Luc S, $\text{LucWT}_{\textit{E. coli}}$ and $\text{LucWT}_{\text{Firefly}}$ were done in parallel, these were analyzed both by Luc expression category as well as by replicate number, with pACYC177 datasets serving as the reference strain. The other experiments were only analyzed in terms of the Luc expression category, also with pACYC177 datasets serving as the reference strain.

5.2.5 Fitting the relationship between Luc expression and σ^{32} activity

The following molecular mechanism was used to understand the relationship between Luc expression, using enzymatic activity as a proxy for the concentration of enzyme, and σ^{32} activity, defined as the percentage of sequencing reads mapped to genes regulated by σ^{32}:

$$\sigma^{32} \cdot \text{DnaK/DnaJ} + \text{Rpol} \overset{\text{Luc}}{\rightleftharpoons} \sigma^{32} \cdot \text{Rpol} + \text{DnaK/DnaJ} \cdot \text{Luc} \tag{5.1}$$

This binding relationship was expressed in terms of a single-site binding model, corresponding to the known molecular mechanism. Here, the two experimental observables are the percentage of ribosome footprints which map to the known members of the σ^{32} regulon, which we denote as $\%\text{FP}_{\sigma^{32}}$, and the Luc enzymatic activity, a proxy for the Luc cellular concentration [Luc]. These observables are related to one another *via* the two baselines for the apparent σ^{32} activity, when [Luc] is zero and when [Luc] is infinite, as well as an equilibrium constant K. To perform nonlinear least-squares fitting with error in both the independent and dependent variables, the standard errors were provided for both the ribosome profiling and luciferase activity measurements, and fits were performed using Orear’s effective variance method in Gnuplot v. 5.0.5.

$$\%\text{FP}_{\sigma^{32}} = \frac{\sigma^{32}{}_{[\text{Luc}]=0} + \sigma^{32}{}_{[\text{Luc}]=\infty} \cdot K \cdot [\text{Luc}]}{1 + K \cdot [\text{Luc}]} \tag{5.2}$$

5.3 Results

5.3.1 Expression of Luc induces the σ^{32} regulon and represses the σ^{38} regulon

It was previously suggested that synonymous changes to the Luc mRNA which slow down translation can improve the folding of the Luc protein, and that changes which speed up translation worsen folding [Spencer et al., 2012, Yu et al., 2015]. For a given amount of protein expression, we predict that these synonymous changes, and their corresponding effects on the folding of the Luc protein, would manifest themselves through differential induction of the chaperones *via* the σ^{32} system. We therefore used ribosome profiling to compare total protein synthesis as a function of Luc expression. Unlike mRNA-seq, which only quantifies the levels of messenger, ribosome profiling measures ribosome occupancy of the mRNA, taking into account differences in post-transcriptional regulation [Li et al., 2014]. Differential translation of chaperone mRNA was analyzed by comparing the total number of sequencing reads mapped to each gene in control cells, with no Luc mRNA, and cells expressing Luc mRNA. Several re-coding strategies were compared: Luc S was re-coded using slow codons, Luc WT using

codons with *E. coli* codons analogous to those in the firefly, and Luc F using fast codons.

As predicted, we observe an induction of the heat shock response following the expression of Luc (Fig. 5.1, p. 192). Importantly, there are significant differences in both the number and magnitude of up-regulated chaperone genes between cells expressing different Luc mRNA. Cells expressing Luc S have only a small number (7) of up-regulated σ^{32} genes, all less than two-fold greater, whereas expression of Luc F turns on many (32) σ^{32} genes, of which roughly half are more than two-fold greater. The cellular response to Luc WT is intermediate (23). Since σ^{32} activity is a direct readout of total unfolded protein in the cell, we conclude that the expression of Luc increases cellular levels of unfolded protein, and that synonymous changes to the Luc mRNA modulate the degree.

Besides the σ^{32} regulon, other genes are affected by Luc expression. In particular, the σ^{38} regulon, which controls the stationary phase response, shows opposite behavior. While Luc S expression turns off few (3) stationary phase genes, expression of either Luc WT (45) or Luc F (42) turns off many genes. However, other significant classes of genes are not differentially regulated. The ribosomal proteins, which constitute 14% of total protein synthesis, are unaffected by Luc expression

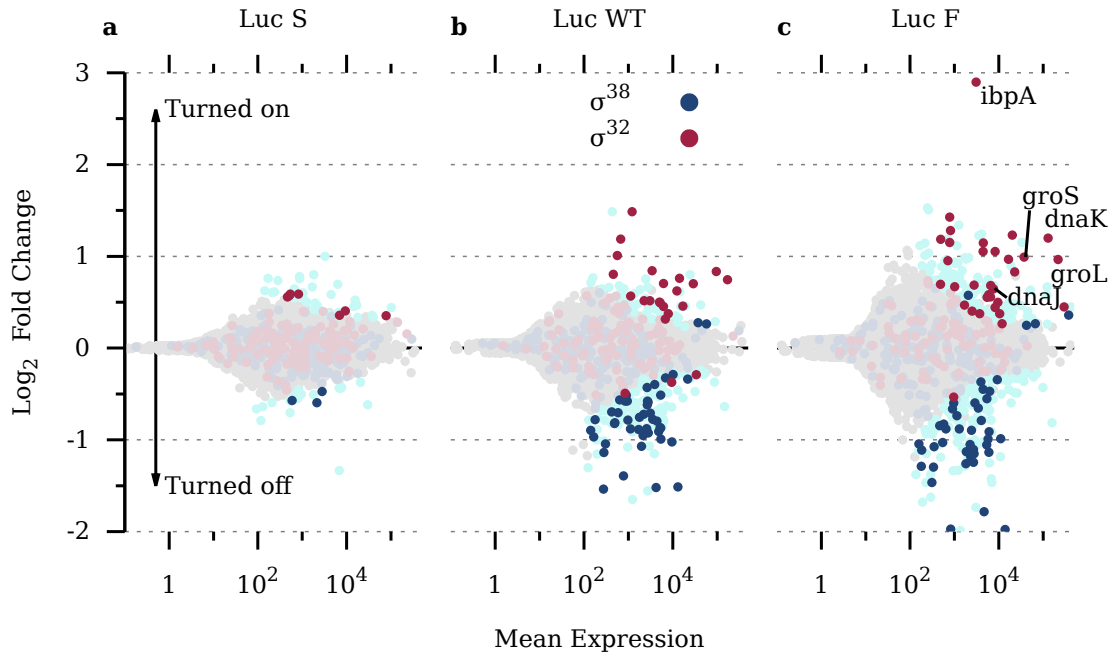


Figure 5.1: Re-coding the Luc mRNA causes major changes to translation levels of the σ^{32} and σ^{38} regulons. Differential mRNA translation analysis of *E. coli* expressing a. Luc S, b. Luc WT, and c. Luc F was performed by comparing the total number of reads mapped to each gene. Dark red and blue circles, and light cyan circles are significantly different, as determined from the reproducibility of three experimental replicates (DeSeq2, $p_{adj} < 0.1$) [Love et al., 2014]). Translation of mRNA whose promoters are targeted by σ^{32} increases following expression of Luc, while translation of mRNA whose promoters are targeted by σ^{38} decreases.

(Fig. 4.12, p. 177), as is β-galactosidase (1.5%), which is under regulation of the same promoter and ribosome binding site sequences as Luc (Fig. 4.7, p. 169).

5.3.2 Localized coding changes induce protein folding stress

Analysis of the induction of the σ^{32} response revealed a relationship to the Luc mRNA sequence. This difference might be caused by perturbation of the translation elongation rate and co-translational folding. If so, this property could be encoded throughout the whole coding sequence, for instance if slow translation promotes co-translational folding in general. Or, it could be localized to certain regions of the gene, indicating that specific structural elements in the Luc protein benefit from slow translation as has been suggested before [Yu et al., 2015]. If the first possibility is true then we predict that hybrid Luc mRNA, containing intermediate number of fast or slow codons, will induce intermediate sigma responses. Conversely, if the differences in chaperone induction are due to sequences at specific locations within the mRNA, then system behavior will depend on the presence or absence of those specific sequences. To test these two possibilities, we analyzed four

additional Luc variants which were assembled by swapping mRNA sequence between Luc S and Luc F. The Luc S/F and Luc F/S variants were split down the middle, with half of the codons re-coded according to the first codon strategy, followed by codons re-coded according to the second strategy. The Luc S_{in}/F and F_{in}/S mRNAs contain just 15 codons (1.5 ribosome widths) of one type, followed by codons of the other type.

We find that the gene expression profiles induced by these four variants fall into the original patterns of either Luc S or Luc F (Fig. 5.2, p. 195). Luc S/F and Luc S_{in}/F, like Luc S, have little to no induction or repression of either the heat shock or stationary phase responses. In contrast, the Luc F/S and F_{in}/S expression profiles are like that of Luc F, with both strong induction of the heat shock response and repression of the stationary phase response. We conclude that the composition of the first 15 codons downstream of the start codon, rather than the totality of the Luc mRNA, is highly predictive of the resulting gene expression changes.

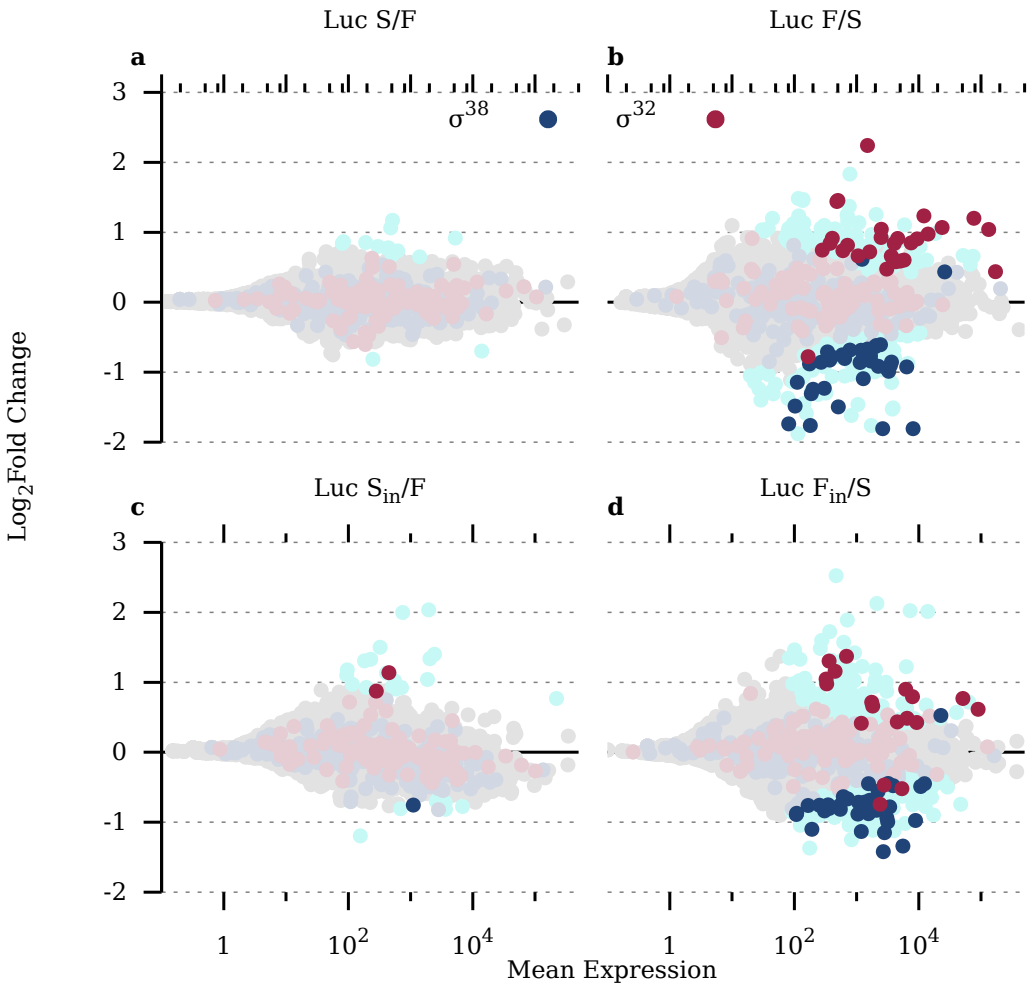


Figure 5.2: Changing the mRNA sequence near the start codon affects the heat shock response. Synonymously re-coded Luc mRNAs were made by swapping either the first 16 (c, d) or 272 (a, b) codons between the fast and slow sequences. The identity of just the first 16 codons has a significant effect on translation. Luc mRNA beginning with slow codons (a, c) have few differentially translated mRNA, whereas those beginning with fast codons (b, d) have many.

5.3.3 Luc expression is regulated by codons immediately downstream of the start codon

We found that changing the 15 codons immediately downstream of the start codon within the Luc mRNA causes changes in unfolded protein levels. It is known that codons near the start codon can affect gene expression by regulating the frequency of translation initiation [Kudla et al., 2009, Pop et al., 2014, Evfratov et al., 2017]. We therefore quantified the effects of codon changes on Luc expression. To compare Luc expression we employed three complementary methods: ribosome profiling, Western blotting and Luciferase activity assays.

Ribosome occupancy of the Luc mRNA shows a clear pattern (Fig. 4.7, p. 169). The three Luc alleles beginning with slow codons (Luc S_{in}/F, Luc S/F, and Luc S) have low ribosome occupancy (0.5-2% of the total), whereas the three alleles beginning with fast codons (Luc F_{in}/S, Luc F/S, and Luc F) have high ribosome occupancy (3-5%). The translation of the wild-type sequence is intermediate (∼2%). These results are confirmed by Western blot and by enzymatic activity (Fig. 5.3, p. 198), with some small differences. Whereas Luc WT has substantially fewer reads than Luc F, Luc F_{in}/S and Luc F/S, the difference is less

pronounced as seen by Western blot. Similarly, the Luc F_{in}/S enzymatic activity levels are comparable to the Luc WT levels, and much lower than Luc F or Luc F/S. We thus observe three clusters of Luc variants with either low, intermediate, or high expression and conclude that the identity of the first 15 codons after the start codon is most predictive of Luc expression levels. To control for the possibility that the observed changes in the σ^{32} response are due to altered levels of sigma factor production, we compared the ribosome profiling data for the sigma factors themselves. We find that the ribosome allocation to *rpoH* mRNA, which encodes σ^{32}, does not change as a function of Luc expression (Fig. 5.4, p. 199).

5.3.4 The σ^{32} response is nonlinear and saturable

We have demonstrated that Luc expression levels depend on the coding of the Luc mRNA. Likewise, the same changes to the mRNA also impact the induction of the σ^{32} regulon and allocation of ribosomes to chaperone mRNA. A comparison of these two measures, Luc enzymatic activity and σ^{32} activity, yields a relationship with saturation characteristics (Fig. 5.5, p. 201). Small increases in Luc levels initially give rise to

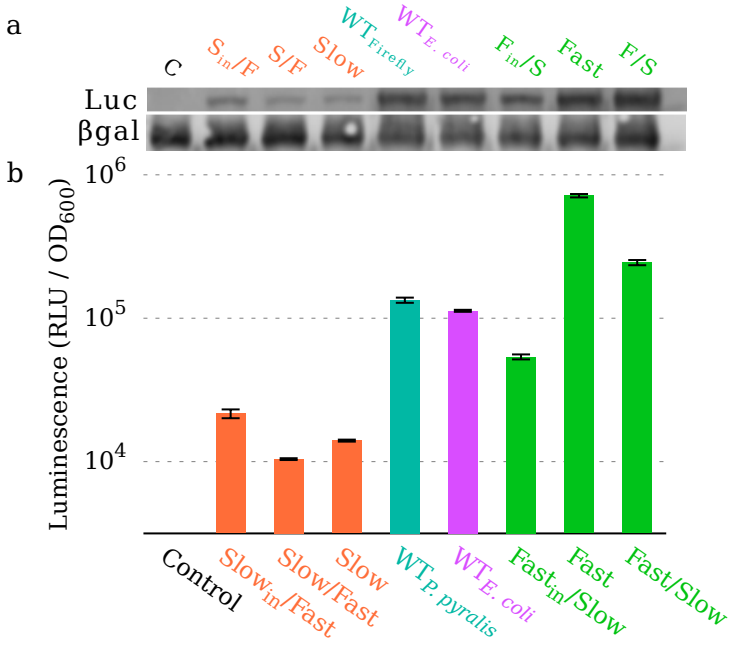


Figure 5.3: Synonymously re-coding the Luc mRNA changes cellular Luc protein levels. Luc protein was quantified using a. Western blot and b. enzymatic activity assays. The mRNA variants beginning with slow codons (Luc S_{in}/F, Luc S/F, Luc S) have low protein levels, while two of the variants beginning with fast codons (Luc F, Luc F/S) have high protein levels. Luc $WT_{Firefly}$, Luc $WT_{\textit{E. coli}}$ and Luc F_{in}/S have intermediate levels of expression.

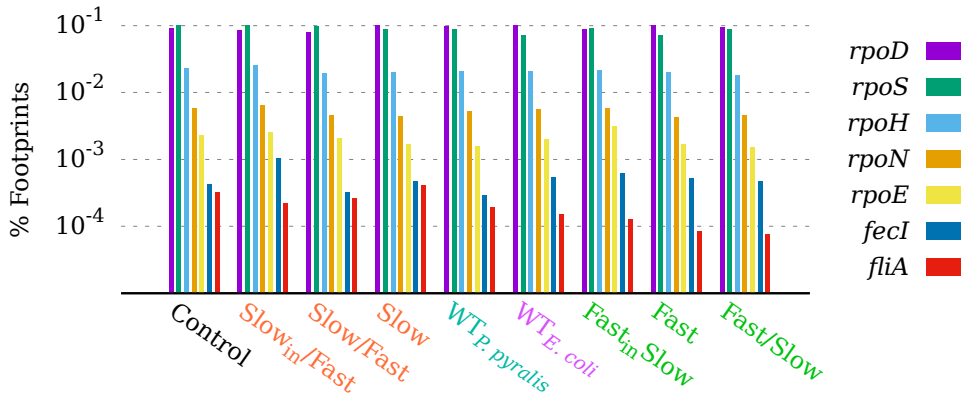


Figure 5.4: Synonymously re-coding the Luc mRNA does not affect *rpoH* ribosome allocation. The production of sigma factors, as measured by ribosome profiling, was determined in relation to the expression of each synonymously re-coded Luc mRNA.

large increases in chaperone allocation, but the effect lessens at higher Luc expression. For example, cells with control plasmid not expressing Luc allocated 5% of ribosomes to the translation of mRNAs which are part of the σ^{32} regulon, while cells expressing Luc WT allocated 6.5%, a 30% increase. Expression of Luc F/S or Luc F further increased allocation to 7%, for a cumulative increase of 40%. However, the amount of Luc F activity required to reach this level of synthesis is more than 600% greater than that of Luc WT.

The σ^{32} transcriptional response has been well characterized, and is known to depend on the competition for binding between RNA polymerase and DnaK/DnaJ [Grigorova et al., 2006, Österberg et al., 2011].

We reasoned that a single-site binding model could be sufficient to describe the observed behavior. In our model, titration of Luc protein into the cytosol shifts the DnaK/DnaJ equilibrium away from binding σ^{32} and towards association with Luc protein (equation 5.1). Under these conditions our model predicts that, in response to increasing amounts of Luc protein, *E. coli* cannot allocate greater than 8% of ribosomes to mRNA under the regulation of σ^{32} (Fig. 5.5). This is equivalent to a maximum induction of 60% from basal levels.

5.3.5 Predicting protein expression with a stochastic translation model

Changing the codons in the Luc CDS (coding sequence) has dramatic effects on the resulting levels of protein production. This effect could be due to changes in mRNA stability, a decrease in the intrinsic initiation rate, or to the occlusion of the start codon or the codons immediately downstream of the start codon. To determine the degree to which this effect is due to occlusion of the start codon by ribosomes, we simulated polysomes for each of the Luc variants and measured the observed translation initiation rates (Fig. 5.6, p. 203), protein production rates (Fig. 5.7, p. 204) and ribosome densities (Fig. 5.8, p. 205) as a function

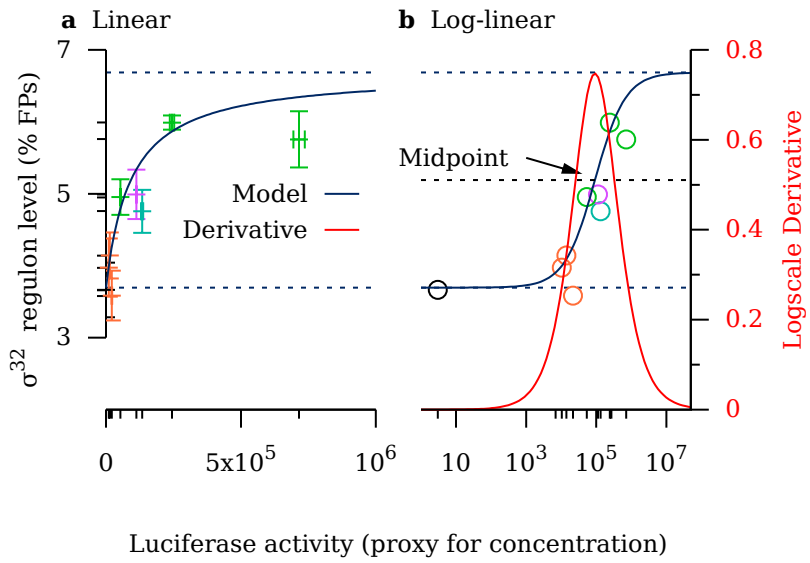


Figure 5.5: The production of Luc protein causes a non-linear, saturable “heat shock” response. Increasing levels of protein (enzymatic activities, x-axis) lead to increasing translation levels of the σ^{32} regulon (ribosome profiling, y-axis). This relationship, modeled using a single-site binding relationship (blue curve), appears to saturate when the Luc mRNA predominantly contains fast codons. The sensitivity of this response is maximal when Luc production is intermediate (red curve). The data are displayed as a. linear-linear and b. log-linear plots.

of the expected translation initiation rate. We find that the appreciable differences in these observables begin to emerge beginning with an expected translation initiation rate of 0.33 s^{-1}. In some cases, the differences are as large as fourfold, indicating that slow codons downstream of the start codon can reduce protein expression. However, the difference is never as great as the twentyfold change observed experimentally, indicating that slow translation is not sufficient to explain the observed protein level variation.

5.4 Discussion

5.4.1 Luc expression modulates the heat shock response

The idea that protein folding efficiency depends on translation rates has received much attention, and Firefly Luciferase has been a popular model protein for a number of reasons. It has multiple domains, one of which can fold while still ribosome-bound [Frydman et al., 1999], and it can be assayed enzymatically. In addition, it folds rapidly in eukaryotic cells yet poorly in *E. coli*, although chaperones are beneficial [Agashe et al., 2004]. Previous experiments showed that the trans-

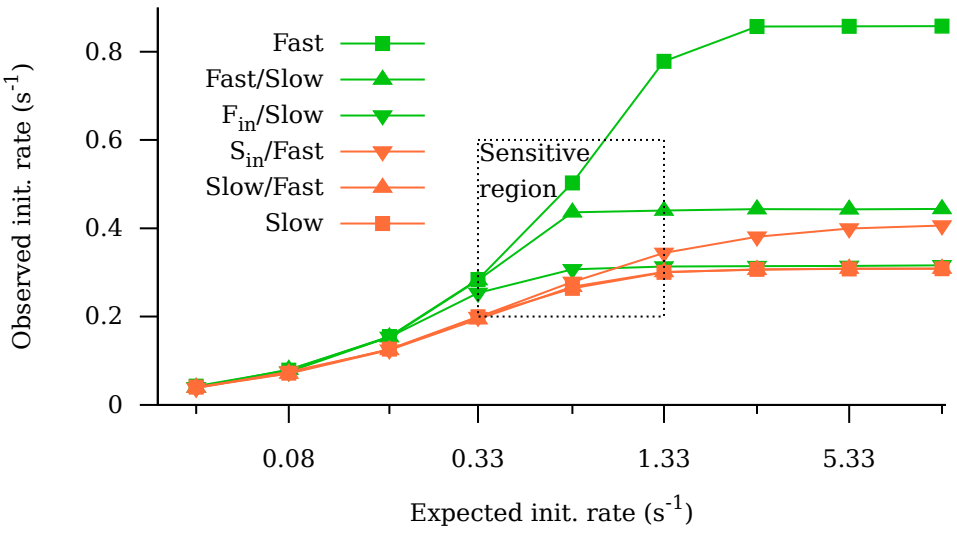


Figure 5.6: Occlusion of the start codon limits translation initiation rates. Polysome simulations were performed by setting slow codons to a translation rate of 5 s^{-1}, fast codons to 20 s^{-1}, and the c-terminal tag to 10 s^{-1}, located in accordance with the re-coding strategy used for the Luc mRNA. Simulations of the six Luc mRNA variants show that, as the initiation rate is allowed to increase, slow codons immediately downstream of the start codon play an increasing role in limiting initiation events. This relationship is particularly sensitive when the initiation rate is between 0.33 and 1.33 events per s.

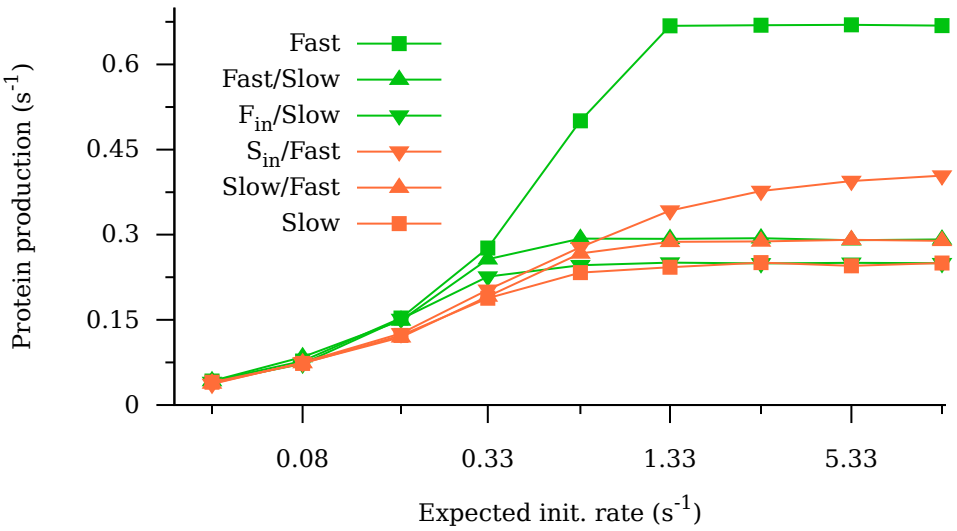


Figure 5.7: Slow codons can limit protein production. Changing the codons in the Luc mRNA is predicted to affect the amount of protein produced over time. This relationship strongly depends on the initiation rate; however the elongation rate is important as well, as evidenced by the higher production rates of the Fast and S_{in}/Fast mRNAs.

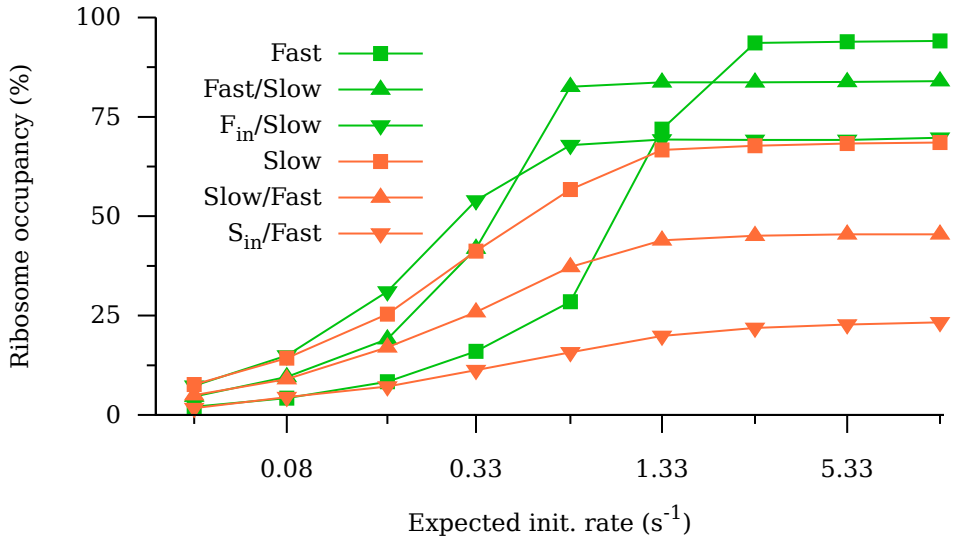


Figure 5.8: Ribosome occupancy of the mRNA depends on the translation rate. Fast codons near the start codon promote loading of ribosomes onto the mRNA, yet they also promote faster completion of protein synthesis, thereby reducing the number of ribosomes simultaneously occupying an mRNA and increasing the spacing between ribosomes. These competing tendencies give rise to mRNA density profiles which are highly sensitive to the initiation rate. In the case of Luc Fast and Fast/Slow, a doubling of the initiation rate can produce a threefold increase in total ribosome occupancy.

lation rate affects the yield of folded Luc protein, both in *E. coli* [Siller et al., 2010, Spencer et al., 2012] and *Neurospora* [Yu et al., 2015]. However, the molecular mechanism determining how, and to what degree, the translation rate affects folding is still unclear. For instance, it could be that the folding of specific structural elements is affected, or that slow translation rates improve folding in general.

We have confirmed that changing the coding of the Luc mRNA does affect the ribosome density, consistent with a change in the translation elongation rate (Ch. 4, p. 136). However, evaluating the Luc protein folding efficiency in live cells is difficult. One approach is to compare the ratio of the overall protein level, such as by gel electrophoresis or Western blot, to the enzymatic activity. Assuming that only folded proteins are enzymatically active, a change in this ratio can indicate changes in protein folding efficiency. However, the cell’s chaperones and proteases act as buffers against unfolded protein, either by promoting folding or degradation of unfolded proteins, meaning that chaperone and protease activity can hide changes to protein folding efficiency. This problem has previously been addressed either by performing experiments in the absence of one or more chaperones [Buhr et al., 2016], or by highly stimulating the production of protein levels, possibly beyond the buffering capacity [Zhang et al., 2009, Spencer et al., 2012]. Unfortunately, these

approaches compromise the estimates of protein folding efficiency under normal conditions.

Rather than bypass the chaperone system, we have used it to our advantage. Since it is known that cells can respond to variable levels of protein folding stress, measuring this response produces an indicator of overall proteostasis. As a proof-of-principle, we have confirmed that ribosome profiling can quantitatively monitor the allocation of ribosomes to mRNAs regulated σ^{32}, reporting the induction of each chaperone, with highly reproducible results. We then developed a physical, quantitative model of how Luc expression activates the chaperone response. To do so also requires an estimate of the cellular levels of Luc protein. We used three different methods to estimate the cellular levels of Luc protein: ribosome profiling, Western blotting and enzymatic assays. Each of these methods comes with trade-offs, but together they present a consistent picture.

The ribosome profiling data measure the allocation of ribosomes to the mRNA in the cell. The data are clearly partitioned into three groups (Fig. 4.7, low, medium and high), as predicted by simulation (Fig. 5.8). However, because the translation rates of the Luc mRNA are not equal, this result does not stand alone. For this reason, we also quantified Luc levels using Western blots with antibodies targetting the C-terminal tag.

The blot confirms that Luc expression falls within three ranges (Fig. 5.3a).

The most sensitive determination of protein levels is the enzymatic assay (Fig. 5.3b). As a whole, we find good agreement between this assay and the other two. Since enzymatic activity only reports on folded, active enzyme, we conclude that Luc folding is minimally impaired. By fitting the data to the single-site binding model we see that the Luc activity assays comport with the σ^{32} response. We conclude that the relationship between Luc production levels and the chaperone response is both nonlinear and saturable.

Remarkably, we observe no direct relationship between coding of the Luc mRNA *per se* and the resulting σ^{32} response. Rather, changing as few as 15 codons at the beginning of the mRNA is sufficient to cause large changes in Luc production, which in turn stimulates the heat shock response. Although this result appears to contradict previous work, it may actually explain several previous results. In our system Luc expression was driven from the P_{lac} promoter on a low-copy plasmid. Even so, the Luc F and Luc F/S mRNAs drove significant amounts of protein synthesis, reaching 85% of the predicted σ^{32} saturation point and 5% of the total ribosome allocation.

Others have expressed Luc using a highly-inducible T7 promoter sys-

tem [Spencer et al., 2012], which likely surpasses the heat shock’s saturation point. Beyond this point, an increase in Luc protein would not be matched by a corresponding increase in DnaK/DnaJ, and the fraction of Luc engaged by chaperones would decrease. One would therefore predict a direct relationship between an increase in total expression levels and unfolded protein levels beyond this point. In fact, those previous experiments clearly showed that the specific activity of Luc protein changes depending on the codons in the Luc mRNA, and we have shown that changing the codons alters the expression level.

5.4.2 Initial codons determine protein production levels

Our use of multiple expression level assays reveals additional differences between expression of the Luc alleles. The ribosome profiling data indicate high levels of Luc F_{in}/S, but the enzymatic activity is intermediate. The high ribosome profiling values could be the result of a bottleneck occurring at the interface of fast and slow codons near the start codon. On the other hand, the enzymatic assays can be interpreted two different ways. One is that enzyme synthesis truly is low, consistent with the effects of a reduced translation initiation rate. The other pos-

sibility is that the translation of this synonymously re-coded Luc mRNA results in poor protein folding. However, this interpretation would be exactly opposite what has been previously reported — that slow translation promotes folding.

Similarly, the fit between Luc enzymatic activity and the heat shock response reveals a potential outlier, Luc S_{in}/F. Whereas the other two Luc mRNAs beginning with slow codons increase the response to about 5.5% of the total footprints, this mRNA actually reduces the chaperone induction, even compared to the control condition with no Luc expression. The reason for this behavior is unclear, but several possibilities come to mind. One is that the Luc proteins translated from this mRNA are extremely well-folded and do not require chaperones. However, this interpretation would suggest that the folding is controlled by fast translation of just the first 16 codons, which would be hard to explain. Furthermore, since the Luc S mRNA was not coded with fast codons yet induces chaperones more, this interpretation would contradict the previous data suggesting that fast translation is bad for folding.

A second possibility is that slow codons are more likely to be mistranslated, causing the insertion of the “wrong” amino acid and producing faulty proteins [Kurland and Gallant, 1996]. If this were the case then the Luc F/S and Luc S/F alleles, which are both coded with ∼50%

slow codons, should have significantly different phenotypes compared to Luc F or Luc S, but they do not.

A third possibility relates to the ribosome density along the S_{in}/F mRNA. We previously showed that the combination of infrequent initiation and fast translation probably depletes the mRNA of ribosomes, leading to mRNA degradation and up-regulation of *ssrA* (Fig. 4.8, p. 171). This response might suppress the induction of chaperones through an unknown mechanism.

Finally, the use of the σ^{32} response to measure Luc-induced protein folding stress assumes that the Luc protein is able to bind to DnaK/DnaJ. If the protein were to exist in an unfolded state which is neither enzymatically active nor binding competent then it would escape detection using this method. In any event, by developing a physical model we are able to show that the behavior of cells expressing Luc S_{in}/F is different.

These large differences in protein expression were unanticipated. In each case, Luc expression is driven by the same low-copy vector backbone, the P_{lac} promoter with full induction using IPTG, and the *lacZ* ribosome binding site. It might, then, be expected that gene copy number, mRNA copy number, and translation initiation frequencies would be identical. Yet we observe a nearly twentyfold change in the amount of protein present, as determined by the enzymatic activity assays. From

our simulations we can ascribe some of this difference to occlusion of the start codon because of slow translation. It may also be the case, however, that the intrinsic translation initiation rate is perturbed, such as by the formation of strong secondary structures in the mRNA which prevent ribosome binding. Finally, it has been shown that 5′ *lacZ* mRNA stability is sensitive to the sequence downstream of the start codon, causing this portion of the mRNA to be preferentially degraded [Petersen, 1991]. Changing the first 15 codons downstream of the start codon has a large effect on expression, and could involve some or all of these mechanisms.

5.4.3 The cellular response to protein overexpression

Ribosome profiling provides a highly sensitive measurement of translation and has been used to characterize protein production at the level of individual proteins beyond those normally detectable by 2D gel electrophoresis. Although the heat shock response up-regulates many genes through σ^{32}, it may be the case that specific chaperones are induced in response to different types of protein folding stress. For example, the core chaperones DnaK/DnaJ and GroEL/GroES are induced roughly

twofold, and the overall change in ribosome allocation increased by 60%. This result is remarkably similar to the effects of temperature shock. It was shown using 2D protein gels that GroEL expression rises about 60% 50 min after a temperature increase, from 28°C to 36°C [Lemaux et al., 1978].

Yet we find that many other σ^{32} targets have different induction levels. Interestingly, the most highly up-regulated protein is IbpA, the inclusion body-binding protein. This protein’s expression is regulated by a temperature-sensitive RNA switch [Gaubig et al., 2011, Kortmann and Narberhaus, 2012], and exposure to heat shock produces a quantitatively similar result to our own — cells grown at 42°C have 8.6 times more IbpA than at 30°C [Mogk et al., 1999], which is similar to the eightfold change we observe. That we observe a similar result in the absence of a temperature change would therefore be unexpected, except that it has previously been noted that IbpA is also highly expressed during exposure to recombinant proteins [Han et al., 2004]. IbpA induction might indicate that the expression of Luc causes an increase in the number of proteins targeted to inclusion bodies.

Unexpectedly, besides the increase in chaperones, we also observe a dramatic down-regulation of genes which comprise the stationary phase response. It has previously been suggested that sigma factors com-

pete with binding to RNA polymerase [Ishihama, 2000, Grigorova et al., 2006, Österberg et al., 2011]. It therefore seems likely that, upon Luc expression, an increase in σ^{32} activity simultaneously prevents σ^{38} from binding RNA polymerase, resulting in a broad reduction in the expression of stationary phase genes. Whether suppression of these genes is directly beneficial or merely a side-effect is unclear.

5.4.4 Implications for protein engineering

Since *E. coli* is a mesophilic enterobacterium, its heat shock response did not evolve under conditions of constant high temperature. Indeed, the *E. coli* proteome is generally maladapted to temperature stress [Dill et al., 2011]. Likewise, the overexpression of foreign proteins, such as Luc, is an artificial scenario. Within the context of typical protein stresses throughout the bacterial lifecycle, we would therefore expect that maximal chaperone induction is modest. *E. coli* is therefore probably not well-suited for overexpressing foreign proteins. Interestingly, directed evolution experiments in *E. coli* have shown that prolonged exposure to elevated temperatures leads to the acquisition of greater temperature resistance and higher chaperone expression, suggesting that the system is tunable in a broader sense [Rudolph et al., 2010].

In the future, ribosome profiling might be used to monitor protein

folding *in vivo*, providing chaperone-specific measurements of induction. For example, one could induce the production of specific amounts of protein within the cell and measure the corresponding changes in σ^{32} activity, perhaps using several combinations of promoters, ribosome binding sites and low or high copy number plasmids. This method could also serve as a means to compare *in vivo* protein folding stress levels, such as in response to changing conditions, or perhaps following changes to a protein’s amino acid sequence.

References

[Adams et al., 1969] Adams, J. M., Jeppesen, P. G., Sanger, F. and Barrell, B. G. (1969). Nucleotide sequence from the coat protein cistron of R17 bacteriophage RNA. Nature *223*, 1009–1014.

[Adzhubei et al., 1996] Adzhubei, A. A., Adzhubei, I. A., Krasheninnikov, I. A. and Neidle, S. (1996). Non-random usage of 'degenerate' codons is related to protein three-dimensional structure. FEBS letters *399*, 78–82.

[Agashe et al., 2004] Agashe, V. R., Guha, S., Chang, H.-C., Genevaux, P., Hayer-Hartl, M., Stemp, M., Georgopoulos, C., Hartl, F. U. and Barral, J. M. (2004). Function of trigger factor and DnaK in multidomain protein folding: increase in yield at the expense of folding speed. Cell *117*, 199–209.

[Air et al., 1976] Air, G. M., Blackburn, E. H., Coulson, A. R., Galibert, F., Sanger, F., Sedat, J. W. and Ziff, E. B. (1976). Gene F of bacteriophage ϕX174. Correlation of nucleotide sequences from the DNA and amino acid sequences from the gene product. Journal of Molecular Biology *107*, 445–458.

[Ames and Hartman, 1963] Ames, B. N. and Hartman, P. E. (1963). The Histidine Operon. Cold Spring Harb Symp Quant Biol *28*, 349–356.

[Anderson, 1969] Anderson, W. F. (1969). The effect of tRNA concentration on the rate of protein synthesis. Proceedings of the National Academy of Sciences of the United States of America *62*, 566–73.

[Andersson et al., 1982] Andersson, D. I., Bohman, K., Isaksson, L. a. and Kurland, C. G. (1982). Translation rates and misreading characteristics of rpsD mutants in Escherichia coli. Molecular & general genetics : MGG *187*, 467–72.

[Andersson et al., 1984] Andersson, S. G., Buckingham, R. H. and Kurland, C. G. (1984). Does codon composition influence ribosome function? The EMBO journal *3*, 91–4.

[Andersson and Kurland, 1990] Andersson, S. G. and Kurland, C. G. (1990). Codon preferences in free-living microorganisms. Microbiological Reviews *54*, 198–210.

[Artieri and Fraser, 2014a] Artieri, C. G. and Fraser, H. B. (2014a). Accounting for biases in riboprofiling data indicates a major role for proline in stalling translation. Genome research *24*, 2011–21.

[Artieri and Fraser, 2014b] Artieri, C. G. and Fraser, H. B. (2014b). Evolution at two levels of gene expression in yeast. Genome research *24*, 411–21.

[Avcilar-Kucukgoze et al., 2016] Avcilar-Kucukgoze, I., Bartholomäus, A., Cordero Varela, J. A., Kaml, R. F.-X., Neubauer, P., Budisa, N. and Ignatova, Z. (2016). Discharging tRNAs: a tug of war between translation and detoxification in Escherichia coli. Nucleic acids research *44*, 8324–34.

[Babbitt et al., 2014] Babbitt, G. A., Alawad, M. A., Schulze, K. V. and Hudson, A. O. (2014). Synonymous codon bias and functional constraint on GC3-related DNA backbone dynamics in the prokaryotic nucleoid. Nucleic acids research *42*, 10915–26.

[Bennetzen and Hall, 1982] Bennetzen, J. L. and Hall, B. D. (1982). Codon selection in yeast. The Journal of biological chemistry *257*, 3026–31.

[Berman et al., 2000] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. and Bourne, P. E. (2000). The protein data bank. Nucleic acids research *28*, 235–242.

[Blundell et al., 1972] Blundell, M., Craig, E. and Kennell, D. (1972). Decay Rates of Different mRNA in E. coli and Models of Decay. Nature New Biology *238*, 46–49.

[Bohman et al., 1984] Bohman, K., Ruusala, T., Jelenc, P. C. and Kurland, C. G. (1984). Kinetic impairment of restrictive streptomycin-resistant ribosomes. Molecular & general genetics : MGG *198*, 90–9.

[Bonekamp et al., 1985] Bonekamp, F., Andersen, H. D., Christensen, T. and Jensen, K. F. (1985). Codon-defined ribosomal pausing in Escherichia coli detected by using the pyrE attenuator to probe the coupling between transcription and translation. Nucleic acids research *13*, 4113–23.

[Bonekamp et al., 1989] Bonekamp, F., Dalbøge, H., Christensen, T. and Jensen, K. F. (1989). Translation rates of individual codons are not correlated with tRNA abundances or with frequencies of utilization in Escherichia coli. Journal of bacteriology *171*, 5812–6.

[Bonekamp and Jensen, 1988] Bonekamp, F. and Jensen, K. F. (1988). The AGG codon is translated slowly in E. coli even at very low expression levels. Nucleic acids research *16*, 3013–24.

[Borg and Ehrenberg, 2015] Borg, A. and Ehrenberg, M. (2015). Determinants of the Rate of mRNA Translocation in Bacterial Protein Synthesis. Journal of Molecular Biology *427*, 1835–1847.

[Boström et al., 1986] Boström, K., Wettesten, M., Borén, J., Bondjers, G., Wiklund, O. and Olofsson, S. O. (1986). Pulse-chase studies of the synthesis and intracellular transport of apolipoprotein B-100 in Hep G2 cells. The Journal of biological chemistry *261*, 13800–6.

[Bremer and Yuan, 1968] Bremer, H. and Yuan, D. (1968). RNA chain growth-rate in Escherichia coli. Journal of Molecular Biology *38*, 163–180.

[Brunak and Engelbrecht, 1996] Brunak, S. and Engelbrecht, J. (1996). Protein structure and the sequential structure of mRNA: alpha-helix and beta-sheet signals at the nucleotide level. Proteins *25*, 237–52.

[Buhr et al., 2016] Buhr, F., Jha, S., Thommen, M., Mittelstaet, J., Kutz, F., Schwalbe, H., Rodnina, M. V. and Komar, A. A. (2016). Synonymous Codons Direct Cotranslational Folding toward Different Protein Conformations. Molecular Cell *61*, 341–351.

[Bulmer, 1987] Bulmer, M. (1987). Coevolution of codon usage and transfer RNA abundance.

[Bulmer, 1991] Bulmer, M. (1991). The selection-mutation-drift theory of synonymous codon usage. Genetics *129*, 897–907.

[Busby et al., 2013] Busby, M. A., Stewart, C., Miller, C. a., Grzeda, K. R. and Marth, G. T. (2013). Scotty: A web tool for designing RNA-Seq

experiments to measure differential gene expression. Bioinformatics *29*, 656–657.

[Chan and Lowe, 2009] Chan, P. P. and Lowe, T. M. (2009). GtRNAdb: a database of transfer RNA genes detected in genomic sequence. Nucleic acids research *37*, D93–7.

[Chaney and Clark, 2015] Chaney, J. L. and Clark, P. L. (2015). Roles for Synonymous Codon Usage in Protein Biogenesis. Annual review of biophysics *44*, 143–166.

[Chaney and Morris, 1978] Chaney, W. G. and Morris, A. J. (1978). Nonuniform size distribution of nascent peptides: The role of messenger RNA. Archives of Biochemistry and Biophysics *191*, 734–741.

[Chaney and Morris, 1979] Chaney, W. G. and Morris, A. J. (1979). Nonuniform size distribution of nascent peptides. Archives of Biochemistry and Biophysics *194*, 283–291.

[Chang and Cohen, 1978] Chang, A. C. and Cohen, S. N. (1978). Construction and characterization of amplifiable multicopy DNA cloning vehicles derived from the P15A cryptic miniplasmid. Journal of bacteriology *134*, 1141–56.

[Charneski and Hurst, 2013] Charneski, C. A. and Hurst, L. D. (2013). Positively Charged Residues Are the Major Determinants of Ribosomal Velocity. PLoS Biology *11*, e1001508.

[Chavancy et al., 1979] Chavancy, G., Chevallier, A., Fournier, A. and Garel, J. P. (1979). Adaptation of iso-tRNA concentration to mRNA codon frequency in the eukaryote cell. Biochimie *61*, 71–8.

[Chavancy and Fournier, 1979] Chavancy, G. and Fournier, A. (1979). Effect of starvation on tRNA synthesis, amino acid pool, tRNA charging levels and aminoacyl-tRNA synthetase activities in the posterior silk gland of Bombyx mori L. Biochimie *61*, 229–243.

[Chavancy and Garel, 1981] Chavancy, G. and Garel, J. P. (1981). Does quantitative tRNA adaptation to codon content in mRNA optimize the ribosomal translation efficiency? Proposal for a translation system model. Biochimie *63*, 187–195.

[Chavancy et al., 1975] Chavancy, G., Garel, J. P. and Daillie, J. (1975). Functional adaptation of aminoacyl-tRNA synthetases to fibroin biosynthesis in the silkgland of Bombyx mori L. FEBS letters *49*, 380–4.

[Chavancy et al., 1981] Chavancy, G., Marbaix, G., Huez, G. and Cleuter, Y. (1981). Effect of tRNA pool balance on rate and uniformity of elongation during translation of fibroin mRNA in a reticulocyte cell-free system. Biochimie *63*, 611–8.

[Chen et al., 2015] Chen, H., Shiroguchi, K., Ge, H. and Xie, X. S. (2015). Genome-wide study of mRNA degradation and transcript elongation in Escherichia coli. Molecular systems biology *11*, 781.

[Chevallier and Garel, 1979] Chevallier, A. and Garel, J. P. (1979). Studies on tRNA adaptation, tRNA turnover, precursor tRNA and tRNA gene distribution in Bombyx mori by using two-dimensional polyacrylamide gel electrophoresis. Biochimie *61*, 245–262.

[Chevallier and Garel, 1982] Chevallier, A. and Garel, J. P. (1982). Differential synthesis rates of tRNA species in the silk gland of Bombyx mori are required to promote tRNA adaptation to silk messages. Eur J Biochem *124*, 477–482.

[Clarke and Clark, 2008] Clarke, T. F. and Clark, P. L. (2008). Rare codons cluster. PloS one *3*, e3412.

[Coffman et al., 1971] Coffman, R. L., Norris, T. E. and Koch, A. L. (1971). Chain elongation rate of messenger and polypeptides in slowly growing Escherichia coli. Journal of molecular biology *60*, 1–19.

[Connolly et al., 1999] Connolly, L., Yura, T. and Gross, C. (1999). Autoregulation of the heat shock response in prokaryotes. In Molecular Chaperones and Folding Catalysts, (Bukau, B., ed.), chapter 2, pp. 13–38. Harwood Academic Publishers Amsterdam.

[Cortazzo et al., 2002] Cortazzo, P., Cerveñansky, C., Marín, M., Reiss, C., Ehrlich, R. and Deana, A. (2002). Silent mutations affect in vivo protein folding in Escherichia coli. Biochemical and biophysical research communications *293*, 537–41.

[Crick, 1966] Crick, F. H. (1966). Codon–anticodon pairing: the wobble hypothesis. Journal of molecular biology *19*, 548–55.

[Crombie et al., 1992] Crombie, T., Swaffield, J. C. and Brown, A. J. P. (1992). Protein folding within the cell is influenced by controlled rates of polypeptide elongation. Journal of Molecular Biology *228*, 7–12.

[Curran and Yarus, 1989] Curran, J. F. and Yarus, M. (1989). Rates of aminoacyl-tRNA selection at 29 sense codons in vivo. Journal of Molecular Biology *209*, 65–77.

[Dahlberg et al., 1973] Dahlberg, A. E., Lund, E. and Kjeldgaard, N. O. (1973). Some effects of antibiotics on bacterial polyribosomes as studied by gel electrophoresis. Journal of molecular biology *78*, 627–636.

[Dalbow and Bremer, 1975] Dalbow, D. G. and Bremer, H. (1975). Metabolic regulation of beta-galactosidase synthesis in Escherichia coli. A test for constitutive ribosome synthesis. The Biochemical journal *150*, 1–8.

[Dalbow and Young, 1975] Dalbow, D. G. and Young, R. (1975). Synthesis time of beta-galactosidase in Escherichia coli B/r as a function of growth rate. The Biochemical journal *150*, 13–20.

[de Marco et al., 2007] de Marco, A., Deuerling, E., Mogk, A., Tomoyasu, T. and Bukau, B. (2007). Chaperone-based procedure to increase yields of soluble recombinant proteins produced in E. coli. BMC Biotechnology *7*, 32.

[Deane and Saunders, 2011] Deane, C. M. and Saunders, R. (2011). The imprint of codons on protein structure. Biotechnology journal *6*, 641–9.

[Dennis and Bremer, 1974] Dennis, P. P. and Bremer, H. (1974). Differential rate of ribosomal protein synthesis in Escherichia coli B/r. Journal of molecular biology *84*, 407–22.

[Dill et al., 2011] Dill, K. A., Ghosh, K. and Schmit, J. D. (2011). Physical limits of cells and proteomes. Proceedings of the National Academy of Sciences of the United States of America *108*, 17876–82.

[Dill et al., 2008] Dill, K. A., Ozkan, S. B., Shell, M. S. and Weikl, T. R. (2008). The protein folding problem. Annual review of biophysics *37*, 289–316.

[Dingwall et al., 1981] Dingwall, C., Lomonossoff, G. P. and Laskey, R. A. (1981). High sequence specificity of micrococcal nuclease. Nucleic acids research *9*, 2659–73.

[Dintzis, 1961] Dintzis, H. M. (1961). Assembly of the peptide chains of hemoglobin. Proceedings of the National Academy of Sciences of the United States of America *47*, 247–61.

[Dittmar et al., 2006] Dittmar, K. A., Goodenbour, J. M. and Pan, T. (2006). Tissue-specific differences in human transfer RNA expression. PLoS genetics *2*, e221.

[Dittmar et al., 2005] Dittmar, K. A., Sørensen, M. A., Elf, J., Ehrenberg, M. and Pan, T. (2005). Selective charging of tRNA isoacceptors induced by amino-acid starvation. EMBO reports *6*, 151–7.

[Doerfel et al., 2012] Doerfel, L. K., Wohlgemuth, I., Kothe, C., Peske, F., Urlaub, H. and Rodnina, M. V. (2012). EF-P Is Essential for Rapid Synthesis of Proteins Containing Consecutive Proline Residues. Science (New York, N.Y.) *339*, 85–88.

[Dong et al., 1995] Dong, H., Nilsson, L. and Kurland, C. G. (1995). Gratuitous overexpression of genes in Escherichia coli leads to growth inhibition and ribosome destruction. Journal of bacteriology *177*, 1497–504.

[Dong et al., 1996] Dong, H., Nilsson, L. and Kurland, C. G. (1996). Covariation of tRNA abundance and codon usage in Escherichia coli at different growth rates. Journal of molecular biology *260*, 649–63.

[dos Reis et al., 2004] dos Reis, M., Savva, R. and Wernisch, L. (2004). Solving the riddle of codon usage preferences: a test for translational selection. Nucleic acids research *32*, 5036–44.

[Efstradiatis et al., 1977] Efstradiatis, A., Maniatis, T. and Kafatos, F. C. (1977). Primary Structure of Rabbit Beta-Globin Messenger-Rna As Determined From Cloned Dna. Cell *10*, 571–585.

[Ehrenberg and Kurland, 1984] Ehrenberg, M. and Kurland, C. G. (1984). Costs of accuracy determined by a maximal growth rate constraint. Quarterly reviews of biophysics *17*, 45–82.

[Ehrenberg et al., 1986] Ehrenberg, M., Kurland, C. G. and Blomberg, C. (1986). Kinetic costs of accuracy in translation. In Accuracy in Molecular Processes, (Kirkwood, T. B. L., Rosenberger, R. F. and Galas, D. J., eds), chapter 11, pp. 329–361. Springer Netherlands Dordrecht.

[Elf et al., 2003] Elf, J., Nilsson, D., Tenson, T. and Ehrenberg, M. (2003). Selective charging of tRNA isoacceptors explains patterns of codon usage. Science (New York, N.Y.) *300*, 1718–22.

[Ellis et al., 2010] Ellis, J. J., Huard, F. P. E., Deane, C. M., Srivastava, S. and Wood, G. R. (2010). Directionality in protein fold prediction. BMC bioinformatics *11*, 172.

[Elton et al., 1976] Elton, R. A., Russell, G. J. and Subak-Sharpe, J. H. (1976). Doublet frequencies and codon weighting in the DNA of Escherichia coli and its phages. Journal of Molecular Evolution *8*, 117–135.

[Engbaek et al., 1973] Engbaek, F., Kjeldgaard, N. O. and Maaloe, O. (1973). Chain growth rate of -galactosidase during exponential growth and amino acid starvation. Journal of molecular biology *75*, 109–18.

[Evans et al., 2008] Evans, M. S., Sander, I. M. and Clark, P. L. (2008). Cotranslational Folding Promotes β-Helix Formation and Avoids Aggregation In Vivo. Journal of Molecular Biology *383*, 683–692.

[Evfratov et al., 2017] Evfratov, S. A., Osterman, I. A., Komarova, E. S., Pogorelskaya, A. M., Rubtsova, M. P., Zatsepin, T. S., Semashko, T. A., Kostryukova, E. S., Mironov, A. A., Burnaev, E., Krymova, E., Gelfand, M. S., Govorun, V. M., Bogdanov, A. A., Sergiev, P. V. and Dontsova, O. A. (2017). Application of sorting and next generation sequencing to study 5´-UTR influence on translation efficiency in Escherichia coli. Nucleic Acids Research *45*, 3487–3502.

[Farabaugh, 1978] Farabaugh, P. J. (1978). Sequence of the lacI gene. Nature *274*, 765–769.

[Farewell and Neidhardt, 1998] Farewell, A. and Neidhardt, F. C. (1998). Effect of temperature on in vivo protein synthetic capacity in Escherichia coli. Journal of bacteriology *180*, 4704–10.

[Fedorov and Baldwin, 1997] Fedorov, A. N. and Baldwin, T. O. (1997). Cotranslational Protein Folding *. Biochemistry *272*, 32715–32718.

[Fedorov et al., 1992] Fedorov, A. N., Friguet, B., Djavadi-Ohaniance, L., Alakhov, Y. B. and Goldberg, M. E. (1992). Folding on the ribosome of Escherichia coli tryptophan synthase ?? subunit nascent chains probed with a conformation-dependent monoclonal antibody. Journal of Molecular Biology *228*, 351–358.

[Fiers et al., 1975] Fiers, W., Contreras, R., Duerinck, F., Haegmean, G., Merregaert, J., Jou, W. M., Raeymakers, A., Volckaert, G., Ysebaert,

M., Van de Kerckhove, J., Nolf, F. and Van Montagu, M. (1975). A-Protein gene of bacteriophage MS2. Nature *256*, 273–278.

[Firnberg et al., 2014] Firnberg, E., Labonte, J. W., Gray, J. J. and Ostermeier, M. (2014). A comprehensive, high-resolution map of a gene's fitness landscape. Molecular biology and evolution *31*, 1581–92.

[Fitch, 1976] Fitch, W. M. (1976). Is there selection against wobble in codon-anticodon pairing? Science (New York, N.Y.) *194*, 1173–4.

[Forchhammer et al., 1972] Forchhammer, J., Jackson, E. N. and Yanofsky, C. (1972). Different half-lives of messenger RNA corresponding to different segments of the tryptophan operon of Escherichia coli. Journal of Molecular Biology *71*, 687–699.

[Forchhammer and Lindahl, 1971] Forchhammer, J. and Lindahl, L. (1971). Growth rate of polypeptide chains as a function of the cell growth rate in a mutant of Escherichia coli 15. Journal of Molecular Biology *55*, 563–568.

[Fournier, 1979] Fournier, A. (1979). Quantitative data on the Bombyx mori L. silkworm: A review. Biochimie *61*, 283–320.

[Frenkel-Morgenstern et al., 2012] Frenkel-Morgenstern, M., Danon, T., Christian, T., Igarashi, T., Cohen, L., Hou, Y.-M. and Jensen, L. J. (2012). Genes adopt non-optimal codon usage to generate cell cycle-dependent oscillations in protein levels. Molecular systems biology *8*, 572.

[Frydman et al., 1999] Frydman, J., Erdjument-Bromage, H., Tempst, P. and Hartl, F. U. (1999). Co-translational domain folding as the structural basis for the rapid de novo folding of firefly luciferase. Nat. Struct. Mol. Biol. *6*, 697–705.

[Fu et al., 2016] Fu, J., Murphy, K. A., Zhou, M., Li, Y. H., Lam, V. H., Tabuloc, C. A., Chiu, J. C. and Liu, Y. (2016). Codon usage affects the structure and function of the Drosophila circadian clock protein PERIOD. Genes & development *30*, 1761–75.

[Galas and Branscomb, 1976] Galas, D. J. and Branscomb, E. W. (1976). Ribosome slowed by mutation to streptomycin resistance. Nature *262*, 617–619.

[Galas et al., 1986] Galas, D. J., Kirkwood, T. B. L. and Rosenberger, R. F. (1986). An introduction to the problem of accuracy. In Accuracy

in Molecular Processes, (Kirkwood, T. B. L., Rosenberger, R. F. and Galas, D. J., eds), chapter 1, pp. 1–16. Champan and Hall New York, NY 1 edition.

[Gamble et al., 2016] Gamble, C. E., Brule, C. E., Dean, K. M., Fields, S. and Grayhack, E. J. (2016). Adjacent Codons Act in Concert to Modulate Translation Efficiency in Yeast. Cell *166*, 679–90.

[Gardin et al., 2014] Gardin, J., Yeasmin, R., Yurovsky, A., Cai, Y., Skiena, S. and Futcher, B. (2014). Measurement of average decoding rates of the 61 sense codons in vivo. eLife *3*, 1–20.

[Garel, 1976] Garel, J.-P. (1976). Quantitative adaptation of isoacceptor tRNAs to mRNA codons of alanine, glycine and serine. Nature *260*, 805–806.

[Garel et al., 1981] Garel, J. P., Chavancy, G., Chevallier, A., Fournier, A., Marbaix, G. and Huez, G. (1981). Adaptation des tRNA et optimisations de la traduction. Reproduction, nutrition, développement *21*, 177–83.

[Garel et al., 1970a] Garel, J. P., Mandel, P., Chavancy, G. and Daillie, J. (1970a). Functional adaptation of tRNAs to fibroin biosynthesis in the silkgland of Bombyx mori L. FEBS letters *7*, 327–329.

[Garel et al., 1970b] Garel, J. P., Virmaux, N. and Mandel, P. (1970b). Adaptation fonctionelle des tRNA à la biosynthèse protéique dans un système cellulaire hautement différencié. I. Mise en évidence dans le cristallin. Bulletin de la Societé de chimie biologique *52*, 987–1006.

[Gaubig et al., 2011] Gaubig, L. C., Waldminghaus, T. and Narberhaus, F. (2011). Multiple layers of control govern expression of the Escherichia coli ibpAB heat-shock operon. Microbiology *157*, 66–76.

[Gausing, 1972] Gausing, K. (1972). Efficiency of protein and messenger RNA synthesis in bacteriophage T4-infected cells of Escherichia coli. Journal of molecular biology *71*, 529–45.

[Gibson et al., 2009] Gibson, D. G., Young, L., Chuang, R.-y., Venter, J. C., Hutchison, C. A. and Smith, H. O. (2009). Enzymatic assembly of DNA molecules up to several hundred kilobases. Nature methods *6*, 343–5.

[Goldstein et al., 1964] Goldstein, A., Goldstein, D. B. and Lowney, L. I. (1964). Protein synthesis at 0°C in Escherichia coli. Journal of Molecular Biology *9*, 213–235.

[Gorini, 1971] Gorini, L. (1971). Ribosomal discrimination of tRNAs. Nature: New biology *234*, 261–4.

[Gorini and Kataja, 1964] Gorini, L. and Kataja, E. (1964). Phenotypic Repair by Streptomycin of Defective Genotypes in E. Coli. Proceedings of the National Academy of Sciences of the United States of America *51*, 487–93.

[Gouy and Grantham, 1980] Gouy, M. and Grantham, R. (1980). Polypeptide elongation and tRNA cycling in Escherichia coli: a dynamic approach. FEBS letters *115*, 151–5.

[Grantham et al., 1980] Grantham, R., Gautier, C., Gouy, M., Mercier, R. and Pavé, A. (1980). Codon catalog usage and the genome hypothesis. Nucleic acids research *8*, r49–r62.

[Grigorova et al., 2006] Grigorova, I. L., Phleger, N. J., Mutalik, V. K. and Gross, C. a. (2006). Insights into transcriptional regulation and sigma competition from an equilibrium model of RNA polymerase binding to DNA. Proceedings of the National Academy of Sciences of the United States of America *103*, 5332–5337.

[Gritsenko et al., 2015] Gritsenko, A. A., Hulsman, M., Reinders, M. J. T. and de Ridder, D. (2015). Unbiased Quantitative Models of Protein Translation Derived from Ribosome Profiling Data. PLoS computational biology *11*, e1004336.

[Grosjean and Fiers, 1982] Grosjean, H. and Fiers, W. (1982). Preferential codon usage in prokaryotic genes: the optimal codon-anticodon interaction energy and the selective codon usage in efficiently expressed genes. Gene *18*, 199–209.

[Guerdoux-Jamet et al., 1997] Guerdoux-Jamet, P., Hénaut, a., Nitschké, P., Risler, J. L. and Danchin, a. (1997). Using codon usage to predict genes origin: is the Escherichia coli outer membrane a patchwork of products from different genomes? DNA research : an international journal for rapid publication of reports on genes and genomes *4*, 257–65.

[Guisbert et al., 2008] Guisbert, E., Yura, T., Rhodius, V. A. and Gross, C. A. (2008). Convergence of Molecular, Modeling, and Systems Approaches for an Understanding of the Escherichia coli Heat Shock Response. Microbiol. Mol. Biol. Rev. *72*, 545–554.

[Gupta and Schlessinger, 1976] Gupta, R. S. and Schlessinger, D. (1976). Coupling of rates of transcription, translation, and messenger ribonucleic acid degradation in streptomycin-dependent mutants of Escherichia coli. Journal of bacteriology *125*, 84–93.

[Gupta et al., 2000] Gupta, S. K., Majumdar, S., Bhattacharya, T. K. and Ghosh, T. C. (2000). Studies on the relationships between the synonymous codon usage and protein secondary structural units. Biochemical and biophysical research communications *269*, 692–6.

[Gutierrez et al., 2013] Gutierrez, E., Shin, B.-S., Woolstenhulme, C. J., Kim, J.-R., Saini, P., Buskirk, A. R. and Dever, T. E. (2013). eIF5A promotes translation of polyproline motifs. Molecular cell *51*, 35–45.

[Guydosh and Green, 2014] Guydosh, N. R. and Green, R. (2014). Dom34 rescues ribosomes in 3' untranslated regions. Cell *156*, 950–62.

[Guydosh and Green, 2017] Guydosh, N. R. and Green, R. (2017). Translation of poly(A) tails leads to precise mRNA cleavage. RNA (New York, N.Y.) *23*, 749–761.

[Haber and Anfinsen, 1962] Haber, E. and Anfinsen, C. B. (1962). Side-chain interactions governing the pairing of half-cystine residues in ribonuclease. The Journal of biological chemistry *237*, 1839–44.

[Han et al., 2004] Han, M.-J., Park, S. J., Park, T. J. and Lee, S. Y. (2004). Roles and applications of small heat shock proteins in the production of recombinant proteins in Escherichia coli. Biotechnology and Bioengineering *88*, 426–436.

[Hansen et al., 2010] Hansen, K. D., Brenner, S. E. and Dudoit, S. (2010). Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucleic acids research *38*, e131.

[Hasegawa et al., 1979] Hasegawa, M., Yasunaga, T. and Miyata, T. (1979). Secondary structure of MS2 phage RNA and bias in code word usage. Nucleic acids research *7*, 2073–9.

[Hatfield et al., 1982] Hatfield, D., Varricchio, F., Rice, M. and Forget, B. G. (1982). The aminoacyl-tRNA population of human reticulocytes. Journal of Biological Chemistry *257*, 3183–3188.

[Heindell et al., 1978] Heindell, H. C., Liu, A., Paddock, G. V., Studnicka, G. M. and Salser, W. A. (1978). The primary sequence of rabbit alpha-globin mRNA. Cell *15*, 43–54.

[Holtkamp et al., 2015] Holtkamp, W., Kokic, G., Jager, M., Mittelstaet, J., Komar, A. A. and Rodnina, M. V. (2015). Cotranslational protein folding on the ribosome monitored in real time. Science *350*, 1104–1107.

[Hopfield, 1974] Hopfield, J. J. (1974). Kinetic proofreading: a new mechanism for reducing errors in biosynthetic processes requiring high specificity. Proceedings of the National Academy of Sciences of the United States of America *71*, 4135–9.

[Hörz and Altenburger, 1981] Hörz, W. and Altenburger, W. (1981). Sequence specific cleavage of DNA by micrococcal nuclease. Nucleic acids research *9*, 2643–58.

[Hussmann et al., 2015] Hussmann, J. A., Patchett, S., Johnson, A., Sawyer, S. and Press, W. H. (2015). Understanding Biases in Ribosome Profiling Experiments Reveals Signatures of Translation Dynamics in Yeast. PLoS Genetics *11*, 1–25.

[Ikemura, 1981a] Ikemura, T. (1981a). Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. Journal of molecular biology *151*, 389–409.

[Ikemura, 1981b] Ikemura, T. (1981b). Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes. Journal of molecular biology *146*, 1–21.

[Ikemura, 1982] Ikemura, T. (1982). Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and Escherichia coli with reference to the abundance of isoaccepting transfer R. Journal of molecular biology *158*, 573–97.

[Ikemura, 1985] Ikemura, T. (1985). Codon usage and tRNA content in unicellular and multicellular organisms. Molecular biology and evolution *2*, 13–34.

[Ikemura and Ozeki, 1983] Ikemura, T. and Ozeki, H. (1983). Codon usage and transfer RNA contents: organism-specific codon-choice patterns in reference to the isoacceptor contents. Cold Spring Harbor symposia on quantitative biology *47 Pt 2*, 1087–97.

[Ingolia et al., 2011] Ingolia, N., Lareau, L. and Weissman, J. (2011). Ribosome Profiling of Mouse Embryonic Stem Cells Reveals the Complexity and Dynamics of Mammalian Proteomes. Cell *147*, 789–802.

[Ingolia et al., 2012] Ingolia, N. T., Brar, G. A., Rouskin, S., McGeachy, A. M. and Weissman, J. S. (2012). The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. Nature protocols *7*, 1534–50.

[Ingolia et al., 2013] Ingolia, N. T., Brar, G. a., Rouskin, S., McGeachy, A. M. and Weissman, J. S. (2013). Genome-wide annotation and quantitation of translation by ribosome profiling. Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.] *Chapter 4*, Unit4.18.

[Ingolia et al., 2014] Ingolia, N. T., Brar, G. A., Stern-Ginossar, N., Harris, M. S., Talhouarne, G. J. S., Jackson, S. E., Wills, M. R. and Weissman, J. S. (2014). Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. Cell reports *8*, 1365–79.

[Ingolia et al., 2009] Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S. and Weissman, J. S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science (New York, N.Y.) *324*, 218–23.

[Iost and Dreyfus, 1995] Iost, I. and Dreyfus, M. (1995). The stability of Escherichia coli lacZ mRNA depends upon the simultaneity of its synthesis and translation. The EMBO journal *14*, 3252–61.

[Ishihama, 2000] Ishihama, A. (2000). Functional modulation of Escherichia coli RNA polymerase. Annual review of microbiology *54*, 499–518.

[Itano, 1957] Itano, H. A. (1957). The Human Hemoglobins: Their Properties and Genetic Control. Advances in Protein Chemistry *12*, 215–268.

[Itano, 1965] Itano, H. A. (1965). The synthesis and structure of normal and abnormal hemoglobins. In Abnormal Hemoglobins in Africa, (Jonxis, J. H. P., ed.), chapter The Synthe, pp. 3–16. Blackwell Scientific Publications Oxford.

[Itano, 1968] Itano, H. A. (1968). The Structure-Rate Hypothesis and the Toll Bridge Analogy. In Structural Chemistry and Molecular Biology, (Rich, A. and Davidson, N., eds), pp. 275–280. W. H. Freeman and Company San Francisco.

[Jorstad and Morris, 1974] Jorstad, C. M. and Morris, D. R. (1974). Polyamine limitation of growth slows the rate of polypeptide chain elongation in Escherichia coli. Journal of bacteriology *119*, 857–60.

[Jou et al., 1971] Jou, W. M., Haegeman, G. and Fiers, W. (1971). Studies on the bacteriophage MS2. Nucleotide fragments from the coat protein cistron. FEBS letters *13*, 105–109.

[Kaiser et al., 2011] Kaiser, C. M., Goldman, D. H., Chodera, J. D., Tinoco, I. and Bustamante, C. (2011). The ribosome modulates nascent protein folding. Science (New York, N.Y.) *334*, 1723–7.

[Kato et al., 1990] Kato, M., Nishikawa, K., Uritani, M., Miyazaki, M. and Takemura, S. (1990). The difference in the type of codon-anticodon base pairing at the ribosomal P-site is one of the determinants of the translational rate. Journal of biochemistry *107*, 242–247.

[Keiler, 2015] Keiler, K. C. (2015). Mechanisms of ribosome rescue in bacteria. Nature reviews. Microbiology *13*, 285–97.

[Kennell and Riezman, 1977] Kennell, D. and Riezman, H. (1977). Transcription and translation initiation frequencies of the Escherichia coli lac operon. Journal of molecular biology *114*, 1–21.

[Kepes, 1963] Kepes, A. (1963). Kinetics of Induced Enzyme Synthesis. Determination of the Mean Life of Galactosidase-Specific Messenger Rna. Biochimica et biophysica acta *76*, 293–309.

[Kepes and Beguin, 1966] Kepes, A. and Beguin, S. (1966). Peptide chain initiation and growth in the induced synthesis of beta-galactosidase. Biochimica et biophysica acta *123*, 546–60.

[Khorana, 1968] Khorana, H. G. (1968). Synthetic nucleic acids and the genetic code. JAMA *206*, 1978–82.

[Kiho and Rich, 1964] Kiho, Y. and Rich, A. (1964). Induced Enztyme Formed on Bacterial Polyribosomes. Proceedings of the National Academy of Sciences of the United States of America *51*, 111–118.

[Kim et al., 2009] Kim, S.-Y., Ayyadurai, N., Heo, M.-A., Park, S., Jeong, Y. J. and Lee, S.-G. (2009). Improving the productivity of recombinant protein in Escherichia coli under thermal stress by coexpressing GroELS chaperone system. Journal of microbiology and biotechnology *19*, 72–7.

[Kim et al., 2013] Kim, Y. E., Hipp, M. S., Bracher, A., Hayer-Hartl, M. and Hartl, F. U. (2013). Molecular chaperone functions in protein folding and proteostasis. Annual review of biochemistry *82*, 323–55.

[Kimchi-Sarfaty et al., 2007] Kimchi-Sarfaty, C., Oh, J. M., Kim, I.-W., Sauna, Z. E., Calcagno, A. M., Ambudkar, S. V. and Gottesman, M. M. (2007). A ”silent” polymorphism in the MDR1 gene changes substrate specificity. Science (New York, N.Y.) *315*, 525–8.

[Klopotowski and Wiater, 1965] Klopotowski, T. and Wiater, A. (1965). Synergism of aminotriazole and phosphate on the inhibition of yeast imidazole glycerol phosphate dehydratase. Archives of biochemistry and biophysics *112*, 562–6.

[Komar and Jaenicke, 1995] Komar, A. A. and Jaenicke, R. (1995). Kinetics of translation of γB crystallin and its circularly permutated variant in an in vitro cell-free system: possible relations to codon distribution and protein folding. FEBS Letters *376*, 195–198.

[Komar et al., 1999] Komar, A. A., Lesnik, T. and Reiss, C. (1999). Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation. FEBS letters *462*, 387–91.

[Kortmann and Narberhaus, 2012] Kortmann, J. and Narberhaus, F. (2012). Bacterial RNA thermometers: molecular zippers and switches. Nature reviews. Microbiology *10*, 255–65.

[Kramer and Farabaugh, 2007] Kramer, E. B. and Farabaugh, P. J. (2007). The frequency of translational misreading errors in E. coli is largely determined by tRNA competition. RNA (New York, N.Y.) *13*, 87–96.

[Kramer et al., 2010] Kramer, E. B., Vallabhaneni, H., Mayer, L. M. and Farabaugh, P. J. (2010). A comprehensive analysis of translational missense errors in the yeast Saccharomyces cerevisiae. RNA (New York, N.Y.) *16*, 1797–808.

[Krasheninnikov et al., 1989] Krasheninnikov, I., Komar, A. and Adzhubeĭ, I. (1989). [Role of the code redundancy in determining cotranslational protein folding]. Biokhimiia *54*, 187–200.

[Krüger et al., 1998] Krüger, M. K., Pedersen, S., Hagervall, T. G. and Sørensen, M. A. (1998). The modification of the wobble base of tRNAGlu modulates the translation rate of glutamic acid codons in vivo. Journal of molecular biology *284*, 621–31.

[Kudla et al., 2009] Kudla, G., Murray, A. W., Tollervey, D. and Plotkin, J. B. (2009). Coding-sequence determinants of gene expression in Escherichia coli. Science (New York, N.Y.) *324*, 255–8.

[Kudlicki et al., 1997] Kudlicki, W., Coffman, A., Kramer, G. and Hardesty, B. (1997). Ribosomes and ribosomal RNA as chaperones for folding of proteins. Folding and Design *2*, 101–108.

[Kurland, 1978] Kurland, C. (1978). The role of guanine nucleotides in protein biosynthesis. Biophysical Journal *22*, 373–392.

[Kurland and Gallant, 1996] Kurland, C. and Gallant, J. (1996). Errors of heterologous protein expression. Current Opinion in Biotechnology *7*, 489–493.

[Lacroute and Stent, 1968] Lacroute, F. and Stent, G. S. (1968). Peptide chain growth of -galactosidase in Escherichia coli. Journal of molecular biology *35*, 165–73.

[Langmead and Salzberg, 2012] Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods *9*, 357–359.

[Langmead et al., 2009] Langmead, B., Trapnell, C., Pop, M. and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome biology *10*, R25.

[Lareau et al., 2014] Lareau, L. F., Hite, D. H., Hogan, G. J. and Brown, P. O. (2014). Distinct stages of the translation elongation cycle revealed by sequencing ribosome-protected mRNA fragments. eLife *3*, e01257.

[Lemaux et al., 1978] Lemaux, P. G., Herendeen, S. L., Bloch, P. L. and Neidhardt, F. C. (1978). Transient rates of synthesis of individual polypeptides in E. coli following temperature shifts. Cell *13*, 427–434.

[Levinthal, 1968] Levinthal, C. (1968). Are there pathways for protein folding? Journal de Chimie Physique et de PhysicoChimie Biologique *65*, 44–45.

[Levinthal, 1969] Levinthal, C. (1969). How to fold graciously. Mössbaun Spectroscopy in Biological Systems Proceedings *67*, 22–24.

[Li et al., 2014] Li, G.-W., Burkhardt, D., Gross, C. and Weissman, J. S. (2014). Quantifying Absolute Protein Synthesis Rates Reveals Principles Underlying Allocation of Cellular Resources. Cell *157*, 624–635.

[Li et al., 2012] Li, G.-W., Oh, E. and Weissman, J. S. (2012). The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. Nature *484*, 538–41.

[Link et al., 2008] Link, A. J., Skretas, G., Strauch, E.-M., Chari, N. S. and Georgiou, G. (2008). Efficient production of membrane-integrated and detergent-soluble G protein-coupled receptors in Escherichia coli. Protein science : a publication of the Protein Society *17*, 1857–1863.

[Lipman and Wilbur, 1983] Lipman, D. J. and Wilbur, W. J. (1983). Contextual constraints on synonymous codon choice. Journal of Molecular Biology *163*, 363–376.

[Liu and Song, 2016] Liu, T.-Y. and Song, Y. S. (2016). Prediction of ribosome footprint profile shapes from transcript sequences. Bioinformatics (Oxford, England) *32*, i183–i191.

[Lizardi et al., 1979] Lizardi, P. M., Mahdavi, V., Shields, D. and Candelas, G. (1979). Discontinuous translation of silk fibroin in a reticulocyte cell-free system and in intact silk gland cells. Proceedings of the National Academy of Sciences of the United States of America *76*, 6211–5.

[Love et al., 2014] Love, M. I., Huber, W. and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology *15*, 550.

[Maaløe and Kjeldgaard, 1966] Maaløe, O. and Kjeldgaard, N. O. (1966). Control of Macromolecular Synthesis: A Study of DNA, RNA, and Protein Synthesis in Bacteria. W. A. Benjamin, New York.

[Makino et al., 2011] Makino, T., Skretas, G. and Georgiou, G. (2011). Strain engineering for improved expression of recombinant proteins in bacteria. Microbial cell factories *10*, 32.

[Marin, 2008] Marin, M. (2008). Folding at the rhythm of the rare codon beat. Biotechnology journal *3*, 1047–57.

[Martens et al., 2015] Martens, A. T., Taylor, J. and Hilser, V. J. (2015). Ribosome A and P sites revealed by length analysis of ribosome profiling data. Nucleic Acids Research *43*, 3680–3687.

[Mathews et al., 2007] Mathews, M. B., Sonenberg, N. and Hershey, J. W. B. (2007). Origins and Principles of Translational Control. In Translational Control in Biology and Medicine, (Michael B. Matthews, Nahum Sonenberg and John W. B. Hershey, eds), chapter 1, pp. 1–40. Cold Spring Harbor Laboratory Press Cold Spring Harbor, New York.

[McQuillen et al., 1959] McQuillen, K., Roberts, R. B. and Britten, R. J. (1959). Synthesis of Nascent Protein By Ribosomes in Escherichia Coli. Proceedings of the National Academy of Sciences of the United States of America *45*, 1437–47.

[Médigue et al., 1991] Médigue, C., Rouxel, T., Vigier, P., Hénaut, a. and Danchin, a. (1991). Evidence for horizontal gene transfer in Escherichia coli speciation. Journal of molecular biology *222*, 851–6.

[Menez et al., 2000] Menez, J., Heurgué-Hamard, V. and Buckingham, R. H. (2000). Sequestration of specific tRNA species cognate to the last sense codon of an overproduced gratuitous protein. Nucleic acids research *28*, 4725–32.

[Misra and Reeves, 1985] Misra, R. and Reeves, P. (1985). Intermediates in the synthesis of TolC protein include an incomplete peptide stalled at a rare Arg codon. European journal of biochemistry / FEBS *152*, 151–5.

[Mita et al., 1988] Mita, K., Ichimura, S., Zama, M. and James, T. C. (1988). Specific codon usage pattern and its implications on the secondary structure of silk fibroin mRNA. Journal of Molecular Biology *203*, 917–925.

[Mitarai et al., 2008] Mitarai, N., Sneppen, K. and Pedersen, S. (2008). Ribosome Collisions and Translation Efficiency: Optimization by Codon Usage and mRNA Destabilization. Journal of Molecular Biology *382*, 236–245.

[Miyata et al., 1979] Miyata, T., Hayashida, H., Yasunaga, T. and Hasegawa, M. (1979). The preferential codon usages in variable and constant regions of immunoglobulin genes are quite distinct from each other. Nucleic acids research *7*, 2431–8.

[Moazed and Noller, 1989] Moazed, D. and Noller, H. F. (1989). Intermediate states in the movement of transfer RNA in the ribosome. Nature *342*, 142–8.

[Mogk et al., 1999] Mogk, A., Tomoyasu, T., Goloubinoff, P., Rüdiger, S., Röder, D., Langen, H. and Bukau, B. (1999). Identification of thermolabile Escherichia coli proteins: prevention and reversion of aggregation by DnaK and ClpB. The EMBO journal *18*, 6934–6949.

[Mohammad et al., 2016] Mohammad, F., Woolstenhulme, C., Green, R. and Buskirk, A. (2016). Clarifying the Translational Pausing Landscape in Bacteria by Ribosome Profiling. Cell Reports *14*, 686–694.

[Morris and Hansen, 1973] Morris, D. R. and Hansen, M. T. (1973). Influence of polyamine limitation on the chain growth rates of beta-galactosidase and of its messenger ribonucleic acid. Journal of bacteriology *116*, 588–92.

[Morse and Yanofsky, 1969] Morse, D. E. and Yanofsky, C. (1969). Polarity and the degradation of mRNA. Nature *224*, 329–31.

[Moszer et al., 1999] Moszer, I., Rocha, E. P. and Danchin, A. (1999). Codon usage and lateral gene transfer in Bacillus subtilis. Current opinion in microbiology *2*, 524–8.

[Needleman and Wunsch, 1970] Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology *48*, 443–453.

[Neidhardt et al., 1974] Neidhardt, F. C., Bloch, P. L. and Smith, D. F. (1974). Culture medium for enterobacteria. Journal of bacteriology *119*, 736–47.

[Newman et al., 2016] Newman, Z. R., Young, J. M., Ingolia, N. T. and Barton, G. M. (2016). Differences in codon bias and GC content contribute to the balanced expression of TLR7 and TLR9. Proceedings of the National Academy of Sciences of the United States of America *113*, E1362–71.

[Nicola et al., 1999] Nicola, A. V., Chen, W. and Helenius, A. (1999). Cotranslational folding of an alphavirus capsid protein in the cytosol of living cells. Nature cell biology *1*, 341–345.

[Ninio, 1971] Ninio, J. (1971). Codon-anticodon recognition: the missing triplet hypothesis. Journal of molecular biology *56*, 63–74.

[Ninio, 1973] Ninio, J. (1973). Recognition in nucleic acids and the anticodon families. Progress in nucleic acid research and molecular biology *13*, 301–37.

[Ninio, 1974] Ninio, J. (1974). A semi-quantitative treatment of missense and nonsense suppression in the strA and ram ribosomal mutants of Escherichia coli Evaluation of some molecular parameters of translation in vivo. Journal of Molecular Biology *84*, 297–313.

[Ninio, 1975] Ninio, J. (1975). Kinetic amplification of enzyme discrimination. Biochimie *57*, 587–95.

[Ninio, 2006] Ninio, J. (2006). Multiple stages in codon-anticodon recognition: double-trigger mechanisms and geometric constraints. Biochimie *88*, 963–92.

[Nishihara et al., 2000] Nishihara, K., Kanemori, M. and Yanagi, H. (2000). Overexpression of Trigger Factor Prevents Aggregation of Recombinant Proteins in Escherichia coli Overexpression of Trigger Factor Prevents Aggregation of Recombinant Proteins in Escherichia coli. Applied and Environmental Microbiology *66*, 884–889.

[O'Connor et al., 2013] O'Connor, P. B. F., Li, G.-W., Weissman, J. S., Atkins, J. F. and Baranov, P. V. (2013). rRNA:mRNA pairing alters the length and the symmetry of mRNA-protected fragments in ribosome profiling experiments. Bioinformatics (Oxford, England) *29*, 1488–91.

[Oh et al., 2011] Oh, E., Becker, A. H., Sandikci, A., Huber, D., Chaba, R., Gloge, F., Nichols, R. J., Typas, A., Gross, C. A., Kramer, G., Weissman, J. S. and Bukau, B. (2011). Selective ribosome profiling reveals the cotranslational chaperone action of trigger factor in vivo. Cell *147*, 1295–308.

[Österberg et al., 2011] Österberg, S., Peso-Santos, T. D. and Shingler, V. (2011). Regulation of Alternative Sigma Factor Use. Annual Review of Microbiology *65*, 37–55.

[Palidwor et al., 2010] Palidwor, G. A., Perkins, T. J. and Xia, X. (2010). A general model of codon bias due to GC mutational bias. PloS one *5*, e13431.

[Pape et al., 1998] Pape, T., Wintermeyer, W. and Rodnina, M. V. (1998). Complete kinetic mechanism of elongation factor Tu-dependent binding of aminoacyl-tRNA to the A site of the E. coli ribosome. The EMBO journal *17*, 7490–7.

[Pavlov et al., 2009] Pavlov, M. Y., Watts, R. E., Tan, Z., Cornish, V. W., Ehrenberg, M. and Forster, A. C. (2009). Slow peptide bond formation

by proline and other N-alkylamino acids in translation. Proceedings of the National Academy of Sciences of the United States of America *106*, 50–4.

[Pechmann et al., 2014] Pechmann, S., Chartron, J. W. and Frydman, J. (2014). Local slowdown of translation by nonoptimal codons promotes nascent-chain recognition by SRP in vivo. Nature Structural & Molecular Biology *21*, 1100–1105.

[Pechmann and Frydman, 2013] Pechmann, S. and Frydman, J. (2013). Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. Nature structural & molecular biology *20*, 237–43.

[Pedersen, 1984] Pedersen, S. (1984). Escherichia coli ribosomes translate in vivo with variable rate. The EMBO journal *3*, 2895–8.

[Pedersen et al., 1978] Pedersen, S., Reeh, S. and Friesen, J. D. (1978). Functional mRNA half lives in E. coli. MGG Molecular & General Genetics *166*, 329–336.

[Petersen, 1991] Petersen, C. (1991). Multiple determinants of functional mRNA stability: Sequence alterations at either end of the lacZ gene affect the rate of mRNA inactivation. Journal of Bacteriology *173*, 2167–2172.

[Plotkin and Kudla, 2011] Plotkin, J. B. and Kudla, G. (2011). Synonymous but not the same: the causes and consequences of codon bias. Nature Reviews Genetics *12*, 32–42.

[Pop et al., 2014] Pop, C., Rouskin, S., Ingolia, N. T., Han, L., Phizicky, E. M., Weissman, J. S. and Koller, D. (2014). Causal signals between codon bias, mRNA structure, and the efficiency of translation and elongation. Molecular systems biology *10*, 770.

[Post et al., 1979] Post, L. E., Strycharz, G. D., Nomura, M., Lewis, H. and Dennis, P. P. (1979). Nucleotide sequence of the ribosomal protein gene cluster adjacent to the gene for RNA polymerase subunit beta in Escherichia coli. Proceedings of the National Academy of Sciences *76*, 1697–1701.

[Protzel and Morris, 1974] Protzel, A. and Morris, A. J. (1974). Gel chromatographic analysis of nascent globin chains. Evidence of nonuniform size distribution. The Journal of biological chemistry *249*, 4594–600.

[Purvis et al., 1987] Purvis, I. J., Bettany, A. J. E., Santiago, T. C., Coggins, J. R., Duncan, K., Eason, R. and Brown, A. J. P. (1987). The efficiency of folding of some proteins is increased by controlled rates of translation in vivo. A hypothesis. Journal of Molecular Biology *193*, 413–417.

[Ramos and Laederach, 2014] Ramos, S. B. V. and Laederach, A. (2014). Molecular biology: A second layer of information in {RNA}. Nature *505*, 621–622.

[Randall et al., 1980] Randall, L. L., Josefsson, L. G. and Hardy, S. J. (1980). Novel intermediates in the synthesis of maltose-binding protein in Escherichia coli. European journal of biochemistry / FEBS *107*, 375–379.

[Robinson et al., 1984] Robinson, M., Lilley, R., Little, S., Emtage, J. S., Yarranton, G., Stephens, P., Millican, A., Eaton, M. and Humphreys, G. (1984). Codon usage can affect efficiency of translation of genes in Escherichia coli. Nucleic acids research *12*, 6663–71.

[Rodnina et al., 1996] Rodnina, M. V., Pape, T., Fricke, R., Kuhn, L. and Wintermeyer, W. (1996). Initial binding of the elongation factor Tu.GTP.aminoacyl-tRNA complex preceding codon recognition on the ribosome. The Journal of biological chemistry *271*, 646–52.

[Rodnina and Wintermeyer, 2001] Rodnina, M. V. and Wintermeyer, W. (2001). Fidelity of aminoacyl-tRNA selection on the ribosome: kinetic and structural mechanisms. Annual review of biochemistry *70*, 415–35.

[Rudolph et al., 2010] Rudolph, B., Gebendorfer, K. M., Buchner, J. and Winter, J. (2010). Evolution of Escherichia coli for growth at high temperatures. Journal of Biological Chemistry *285*, 19029–19034.

[Ruusala et al., 1984] Ruusala, T., Andersson, D., Ehrenberg, M. and Kurland, C. G. (1984). Hyper-accurate ribosomes inhibit growth. The EMBO journal *3*, 2575–80.

[Ruusala et al., 1982] Ruusala, T., Ehrenberg, M. and Kurland, C. G. (1982). Is there proofreading during polypeptide synthesis? The EMBO journal *1*, 741–5.

[Sander et al., 2014] Sander, I. M., Chaney, J. L. and Clark, P. L. (2014). Expanding Anfinsen's principle: contributions of synonymous codon

selection to rational protein design. Journal of the American Chemical Society *136*, 858–61.

[Sauna and Kimchi-Sarfaty, 2011] Sauna, Z. E. and Kimchi-Sarfaty, C. (2011). Understanding the contribution of synonymous mutations to human disease. Nature reviews. Genetics *12*, 683–691.

[Saunders and Deane, 2010] Saunders, R. and Deane, C. M. (2010). Synonymous codon usage influences the local protein structure observed. Nucleic acids research *38*, 6719–28.

[Savitzky and Golay, 1964] Savitzky, A. and Golay, M. J. E. (1964). Smoothing and Differentiation of Data by Simplified Least Squares Procedures. Analytical Chemistry *36*, 1627–1639.

[Schleif, 1967] Schleif, R. (1967). Control of production of ribosomal protein. Journal of molecular biology *27*, 41–55.

[Schleif et al., 1973] Schleif, R., Hess, W., Finkelstein, S. and Ellis, D. (1973). Induction kinetics of the L-arabinose operon of Escherichia coli. Journal of bacteriology *115*, 9–14.

[Schrader et al., 2014] Schrader, J. M., Zhou, B., Li, G.-W., Lasker, K., Childers, W. S., Williams, B., Long, T., Crosson, S., McAdams, H. H., Weissman, J. S. and Shapiro, L. (2014). The Coding and Noncoding Architecture of the Caulobacter crescentus Genome. PLoS genetics *10*, e1004463.

[Schwanhäusser et al., 2011] Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W. and Selbach, M. (2011). Global quantification of mammalian gene expression control. Nature *473*, 337–42.

[Sharp and Li, 1987] Sharp, P. M. and Li, W. H. (1987). The codon Adaptation Index–a measure of directional synonymous codon usage bias, and its potential applications. Nucleic acids research *15*, 1281–95.

[Sharp et al., 1986] Sharp, P. M., Tuohy, T. M. and Mosurski, K. R. (1986). Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic acids research *14*, 5125–43.

[Shively, 1974] Shively, J. M. (1974). Inclusion Bodies of Prokaryotes. Annual Review of Microbiology *28*, 167–188.

[Siller et al., 2010] Siller, E., DeZwaan, D. C., Anderson, J. F., Freeman, B. C. and Barral, J. M. (2010). Slowing bacterial translation speed enhances eukaryotic protein folding efficiency. Journal of molecular biology *396*, 1310–18.

[Skretas et al., 2009] Skretas, G., Carroll, S., DeFrees, S., Schwartz, M. F., Johnson, K. F. and Georgiou, G. (2009). Expression of active human sialyltransferase ST6GalNAcI in Escherichia coli. Microbial cell factories *8*, 50.

[Sørensen et al., 1994] Sørensen, M. a., Jensen, K. F. and Pedersen, S. (1994). High concentrations of ppGpp decrease the RNA chain growth rate. Implications for protein synthesis and translational fidelity during amino acid starvation in Escherichia coli. Journal of molecular biology *236*, 441–54.

[Sørensen et al., 1989] Sørensen, M. A., Kurland, C. G. and Pedersen, S. (1989). Codon usage determines translation rate in Escherichia coli. Journal of molecular biology *207*, 365–77.

[Sørensen and Pedersen, 1991] Sørensen, M. A. and Pedersen, S. (1991). Absolute in vivo translation rates of individual codons in Escherichia coli. Journal of Molecular Biology *222*, 265–280.

[Sørensen and Pedersen, 1998] Sørensen, M. A. and Pedersen, S. (1998). Determination of the peptide elongation rate in vivo. Methods in molecular biology (Clifton, N.J.) *77*, 129–42.

[Spanjaard and van Duin, 1988] Spanjaard, R. A. and van Duin, J. (1988). Translation of the sequence AGG-AGG yields 50% ribosomal frameshift. Proceedings of the National Academy of Sciences of the United States of America *85*, 7967–71.

[Spencer and Barral, 2012] Spencer, P. S. and Barral, J. M. (2012). Genetic Code Redundancy and Its Influence on the Encoded Polypeptides. Computational and Structural Biotechnology Journal *1*, 1–8.

[Spencer et al., 2012] Spencer, P. S., Siller, E., Anderson, J. F. and Barral, J. M. (2012). Silent substitutions predictably alter translation elongation rates and protein folding efficiencies. Journal of molecular biology *422*, 328–35.

[Stadler and Fire, 2011] Stadler, M. and Fire, A. (2011). Wobble basepairing slows in vivo translation elongation in metazoans. RNA *17*, 2063–73.

[Steitz, 1969] Steitz, J. A. (1969). Polypeptide Chain Initiation: Nucleotide Sequences of the Three Ribosomal Binding Sites in Bacteriophage R17 RNA. Nature *224*, 957–964.

[Stent, 1964] Stent, G. S. (1964). The Operon: On Its Third Anniversary. Modulation of Transfer RNA Species can Provide a Workable Model of an Operator-Less Operon. Science (New York, N.Y.) *144*, 816–20.

[Stergachis et al., 2013] Stergachis, A. B., Haugen, E., Shafer, A., Fu, W., Vernot, B., Reynolds, A., Raubitschek, A., Ziegler, S., LeProust, E. M., Akey, J. M. and Stamatoyannopoulos, J. A. (2013). Exonic transcription factor binding directs codon choice and affects protein evolution. Science (New York, N.Y.) *342*, 1367–72.

[Subramaniam et al., 2014] Subramaniam, A., Zid, B. and O'Shea, E. (2014). An Integrated Approach Reveals Regulatory Controls on Bacterial Translation Elongation. Cell *159*, 1200–1211.

[Suzuki and Shimodaira, 2006] Suzuki, R. and Shimodaira, H. (2006). Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics (Oxford, England) *22*, 1540–2.

[Takyar et al., 2005] Takyar, S., Hickerson, R. P. and Noller, H. F. (2005). mRNA helicase activity of the ribosome. Cell *120*, 49–58.

[Talkad et al., 1976] Talkad, V., Schneider, E. and Kennell, D. (1976). Evidence for variable rates of ribosome movement in Escherichia coli. Journal of molecular biology *104*, 299–303.

[Tashiro et al., 1968] Tashiro, Y., Morimoto, T., Matsuura, S. and Nagata, S. (1968). STUDIES ON THE POSTERIOR SILK GLAND OF THE SILKWORM, BOMBYX MORI: I. Growth of Posterior Silk Gland Cells and Biosynthesis of Fibroin During the Fifth Larval Instar. The Journal of Cell Biology *38*, 574–588.

[Thanaraj and Argos, 1996a] Thanaraj, T. A. and Argos, P. (1996a). Ribosome-mediated translational pause and protein domain organization. Protein science : a publication of the Protein Society *5*, 1594–1612.

[Thanaraj and Argos, 1996b] Thanaraj, T. A. and Argos, P. (1996b). Protein secondary structural types are differentially coded on messenger RNA. Protein science : a publication of the Protein Society *5*, 1973–1983.

[Thompson and Stone, 1977] Thompson, R. C. and Stone, P. J. (1977). Proofreading of the codon-anticodon interaction on ribosomes. Proceedings of the National Academy of Sciences of the United States of America *74*, 198–202.

[Thurman et al., 2012] Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., Sheffield, N. C., Stergachis, A. B., Wang, H., Vernot, B., Garg, K., John, S., Sandstrom, R., Bates, D., Boatman, L., Canfield, T. K., Diegel, M., Dunn, D., Ebersol, A. K., Frum, T., Giste, E., Johnson, A. K., Johnson, E. M., Kutyavin, T., Lajoie, B., Lee, B.-K., Lee, K., London, D., Lotakis, D., Neph, S., Neri, F., Nguyen, E. D., Qu, H., Reynolds, A. P., Roach, V., Safi, A., Sanchez, M. E., Sanyal, A., Shafer, A., Simon, J. M., Song, L., Vong, S., Weaver, M., Yan, Y., Zhang, Z., Zhang, Z., Lenhard, B., Tewari, M., Dorschner, M. O., Hansen, R. S., Navas, P. a., Stamatoyannopoulos, G., Iyer, V. R., Lieb, J. D., Sunyaev, S. R., Akey, J. M., Sabo, P. J., Kaul, R., Furey, T. S., Dekker, J., Crawford, G. E. and Stamatoyannopoulos, J. a. (2012). The accessible chromatin landscape of the human genome. Nature *489*, 75–82.

[Trovato and O'Brien, 2016] Trovato, F. and O'Brien, E. P. (2016). Insights into Cotranslational Nascent Protein Behavior from Computer Simulations. Annual review of biophysics *45*, 345–69.

[Tsai et al., 2008] Tsai, C.-J., Sauna, Z. E., Kimchi-Sarfaty, C., Ambudkar, S. V., Gottesman, M. M. and Nussinov, R. (2008). Synonymous mutations and ribosome stalling can lead to altered folding pathways and distinct minima. Journal of molecular biology *383*, 281–91.

[Tubulekas and Hughes, 1993] Tubulekas, I. and Hughes, D. (1993). Suppression of rpsL phenotypes by tuf mutations reveals a unique relationship between translation elongation and growth rate. Molecular microbiology *7*, 275–84.

[Ude et al., 2012] Ude, S., Lassak, J., Starosta, A. L., Kraxenberger, T., Wilson, D. N. and Jung, K. (2012). Translation Elongation Factor EF-P Alleviates Ribosome Stalling at Polyproline Stretches. Science (New York, N.Y.) *339*, 82–85.

[Varenne et al., 1989] Varenne, S., Baty, D., Verheij, H., Shire, D. and Lazdunski, C. (1989). The maximum rate of gene expression is dependent in the downstream context of unfavourable codons. Biochimie *71*, 1221–1229.

[Varenne et al., 1984] Varenne, S., Buc, J., Lloubes, R. and Lazdunski, C. (1984). Translation is a non-uniform process. Effect of tRNA availability on the rate of elongation of nascent polypeptide chains. Journal of molecular biology *180*, 549–76.

[Varenne et al., 1982] Varenne, S., Knibiehler, M., Cavard, D., Morlon, J. and Lazdunski, C. (1982). Variable rate of polypeptide chain elongation for colicins A, E2 and E3. Journal of molecular biology *159*, 57–70.

[Varenne and Lazdunski, 1986] Varenne, S. and Lazdunski, C. (1986). Effect of distribution of unfavourable codons on the maximum rate of gene expression by an heterologous organism. Journal of Theoretical Biology *120*, 99–110.

[von der Haar, 2012] von der Haar, T. (2012). Mathematical and Computational Modelling of Ribosomal Movement and Protein Synthesis: an overview. Computational and structural biotechnology journal *1*, 1–7.

[Wang et al., 2016a] Wang, C., Han, B., Zhou, R. and Zhuang, X. (2016a). Real-Time Imaging of Translation on Single mRNA Transcripts in Live Cells. Cell *165*, 990–1001.

[Wang et al., 2016b] Wang, H., McManus, J. and Kingsford, C. (2016b). Accurate Recovery of Ribosome Positions Reveals Slow Translation of Wobble-Pairing Codons in Yeast. In Methods in molecular biology (Clifton, N.J.), (Singh, M., ed.), vol. 1358, of Lecture Notes in Computer Science pp. 37–52. Springer International Publishing Cham.

[Weinberg et al., 2016] Weinberg, D. E., Shah, P., Eichhorn, S. W., Hussmann, J. A., Plotkin, J. B. and Bartel, D. P. (2016). Improved Ribosome-Footprint and mRNA Measurements Provide Insights into Dynamics and Regulation of Yeast Translation. Cell Reports *14*, 1787–1799.

[Wolin and Walter, 1988] Wolin, S. L. and Walter, P. (1988). Ribosome pausing and stacking during translation of a eukaryotic mRNA. The EMBO journal *7*, 3559–69.

[Woolstenhulme et al., 2015] Woolstenhulme, C., Guydosh, N., Green, R. and Buskirk, A. (2015). High-Precision Analysis of Translational Pausing by Ribosome Profiling in Bacteria Lacking EFP. Cell Reports *11*, 13–21.

[Xie and Ding, 1998] Xie, T. and Ding, D. (1998). The relationship between synonymous codon usage and protein structure. FEBS letters *434*, 93–6.

[Xu et al., 2013] Xu, Y., Ma, P., Shah, P., Rokas, A., Liu, Y. and Johnson, C. H. (2013). Non-optimal codon usage is a mechanism to achieve circadian clock conditionality. TL - 495. Nature *495 VN* -, 116–120.

[Yang et al., 2014] Yang, J.-R., Chen, X. and Zhang, J. (2014). Codon-by-Codon Modulation of Translational Speed and Accuracy Via mRNA Folding. PLoS Biology *12*, e1001910.

[Yanofsky, 1981] Yanofsky, C. (1981). Attenuation in the control of expression of bacterial operons. Nature *289*, 751–758.

[Yarus and Smith, 1995] Yarus, M. and Smith, D. (1995). tRNA on the Ribosome: a Waggle Theory. In tRNA: Structure, Biosynthesis, and Function, (Söll, D. and RajBhandary, U. L., eds), chapter tRNA on th, pp. 443–469. American Society for Microbiology Press Washington, D.C.

[Yegian et al., 1966] Yegian, C. D., Stent, G. S. and Martin, E. M. (1966). Intracellular condition of Escherichia coli transfer RNA. Proceedings of the National Academy of Sciences of the United States of America *55*, 839–46.

[Young and Bremer, 1976] Young, R. and Bremer, H. (1976). Polypeptide-chain-elongation rate in Escherichia coli B/r as a function of growth rate. The Biochemical journal *160*, 185–94.

[Yu et al., 2015] Yu, C.-H., Dang, Y., Zhou, Z., Wu, C., Zhao, F., Sachs, M. S. and Liu, Y. (2015). Codon Usage Influences the Local Rate of Translation Elongation to Regulate Co-translational Protein Folding. Molecular Cell *59*, 744–754.

[Zaher and Green, 2009] Zaher, H. S. and Green, R. (2009). Fidelity at the Molecular Level: Lessons from Protein Synthesis. Cell *136*, 746–762.

[Zengel et al., 1977] Zengel, J. M., Young, R., Dennis, P. P. and Nomura, M. (1977). Role of ribosomal protein S12 in peptide chain elongation: analysis of pleiotropic, streptomycin-resistant mutants of Escherichia coli. Journal of bacteriology *129*, 1320–9.

[Zhang et al., 2009] Zhang, G., Hubalewska, M. and Ignatova, Z. (2009). Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. Nature structural & molecular biology *16*, 274–80.

[Zhang and Qian, 2011] Zhang, X. and Qian, S. B. (2011). Chaperone-mediated hierarchical control in targeting misfolded proteins to aggresomes. Mol Biol Cell *22*, 3277–3288.

[Zhou et al., 2000] Zhou, C. Z., Confalonieri, F., Medina, N., Zivanovic, Y., Esnault, C., Yang, T., Jacquet, M., Janin, J., Duguet, M., Perasso, R. and Li, Z. G. (2000). Fine organization of Bombyx mori fibroin heavy chain gene. Nucleic acids research *28*, 2413–9.

[Zhou et al., 2013] Zhou, M., Guo, J., Cha, J., Chae, M., Chen, S., Barral, J. M., Sachs, M. S. and Liu, Y. (2013). Non-optimal codon usage affects expression, structure and function of clock protein FRQ. Nature *495*, 111–5.

[Zipser and Perrin, 1963] Zipser, D. and Perrin, D. (1963). Complementation on Ribosomes. Cold Spring Harbor Symposia on Quantitative Biology *28*, 533–537.

Vita

Andrew Ted Martens was born December 31, 1986 in Arlington Co., Virginia. He grew up in both France and the United States. After graduating from *New Trier High School*, in 2004, he attended *Davidson College*, where he majored in biology and graduated with honors in 2008.

Martens' first research experience was with Dr. Sean Crosson at the *University of Chicago*, during the summer of 2006, where he studied the bacterium *Caulobacter crescentus* and the mechanisms controlling its energy uptake and metabolism. While at Davidson he was a

member of the 2007 iGEM synthetic biology team, led by Dr. A. Malcolm Campbell and Dr. Laurie Heyer, where he worked on the development of bacterial computers to solve the “Hamiltonian Path” problem. He simultaneously worked on an honors thesis, implementing the two-spatula burnt-pancake problem in *E. coli* with the use of the tryptophan anti-terminator transcriptional system. He was also a member of the Bernard Society of Mathematics and the βββ Biology Society.

Martens spent the following year in Paris, France, where he worked in Dr. Antoine Danchin’s laboratory at *Institut Pasteur*. This work was the basis for his Masters degree, awarded in 2009 by *Université Pierre et Marie Curie*. His curriculum included a bioinformatics course on genome analysis and a marine microbiology elective. His Masters thesis was on the metabolic mechanisms of serine toxicity and degradation in *E. coli*. In 2010, Martens enrolled in the *Johns Hopkins University* Department of Biology’s CMDB graduate program, and he joined Dr. Vincent Hilser’s laboratory in 2011.

While at *Johns Hopkins*, Martens actively promoted science education. He made numerous visits to the *Baltimore Talent Development High School*, where he and other students brought hands-on science demonstrations for high school biology classes. During his second year, he taught introductory laboratory courses in biochemistry and cell biol-

ogy for *Johns Hopkins* undergraduates. The following year he mentored four rotation students. In addition, he also served for three years as teaching assistant, and then as course instructor, helping to develop a new computational biology curriculum for the CMDB graduate program.